链滴

微软为什么和联通有仇

作者：localvar
原文链接：https://ld246.com/article/1524893471325
来源网站：链滴

网上流传着一个笑话，说微软和联通有仇，内容大致如下：如果你的电脑操作系统是WIN2000或WINP，那么：

1. 在桌面上点右键，选择新建 — 文本文档；
2. 打开"新建 文本文档"，录入　移动两字后存储后关掉
3. 重新打开"新建 文本文档"，看到什么了？是不是刚刚录入的"移动"两字？
4. 把　移动分别换成电信和网通，重复1--3步，是不是也都没什么问题？
5. 现在我们拿　联通来试试，重复1--3步，你会发现刚刚录入的联通两字不见了，取而代之是个烧焦手机电池(一个符号)。 看来微软确实跟联通有仇呀！

笑话当然是笑话，不能当真。但为什么会这样呢？是微软的bug吗？确实有点像，不过——微软是世顶级的软件公司，记事本则有可能是windows中最简单应用程序，说这是bug未免有点不合情理吧？

好了，既然把自己的主观臆断否定了，就让我们踏上寻找事实真相的艰苦历程吧:)。

不知你注意过没有，记事本的打开、保存对话框比普通的文件对话框多一个编码选项，可以通过它指文件的编码是UNICODE、ANSI还是UTF8。"喔，我知道了"，你可能会说，"这肯定是Windows API IsTextUnicode惹的祸。因为文本文件本身不保存编码信息，所以记事本打开文件时就要调用IsTextUnicode来判断文件的编码。而IsTextUnicode是根据文本的内容猜测其编码，所以肯定是它猜错编码格了。想想 '联通'只有两个字，这样的错误有情可原，OK了，问题解决了"。

说实话，一开始我也是这么想的，但后来发现，我犯了两个错误：

1. IsTextUnicode并没有猜错，不信你可以检查一下　IsTextUnicode("联通", 4, NULL)的返回值。
2. 记事本有可能保存编码信息，这个后面再说。

原来，记事本除了判断编码是不是UNICODE以外，还要判断它是不是UTF8。"联通"两个字的代码是(节顺序从低到高)：C1 AA CD A8，转换为二进制是：11000001 10101010 11001101 10101000。照UTF8编码方案(详情请见http://www.cis.ohio-state.edu/htbin/rfc/rfc2279.html)：

- 0000-007F之间的字符不做转换
- 0080-07FF之间的编码为110xxxxx 10xxxxxx
- 0800-FFFF之间的编码为1110xxxx 10xxxxxx 10xxxxxx

不难发现，"联通"的编码符合第二种情况，所以记事本把它判定为UTF8编码，而对其进行解码后，将成00000000 01101010 00000011 01101000。注意：前两个字节解码后并不在0080--07FF之间，以被认为是错误的值，忽略了。后面两个字节经过调整字节顺序后，将变为16进制的0x0368，也就那块烧毁的电池了(取决于所使用的字体)。

PS:

1. 如果你保存文件时，指定使用除　ANSI以外的编码，记事本将用文件开头的几个字节保存文件编，UNICODE对应0xFEFF，UNICODE BIG ENDIAN对应0xFFFE，UTF8对应0xBFBBEF。这几个字节称为BOM(byte order mark, 字节顺序标记)。如果文件有BOM，记事本直接使用它判断编码，否则就根据文件内容判断编码。
2. 分析的过程中我用　UltraEdit来查看文件的16进制内容，但它会自动进行编码转换并给文件加上个BOM，导致看到的和实际不符(文件4字节，到了UltraEdit中就成了6字节)，让我走了一些弯路。