Some files come up strange in Notepad

devblogs.microsoft.com/oldnewthing/20040324-00

March 24, 2004

Raymond Chen

David Cumps discovered that certain text files come up strange in Notepad. The reason is that Notepad has to edit files in a variety of encodings, and when its back against the wall, sometimes it's forced to guess. Here's the file "Hello" in various encodings:

```
48 65 6C 6C 6F
```

This is the traditional ANSI encoding.

```
48 00 65 00 6C 00 6C 00 6F 00
```

This is the Unicode (little-endian) encoding with no BOM.

```
FF FE 48 00 65 00 6C 00 6C 00 6F 00
```

This is the Unicode (little-endian) encoding with BOM. The BOM (FF FE) serves two purposes: First, it tags the file as a Unicode document, and second, the order in which the two bytes appear indicate that the file is little-endian.

```
00 48 00 65 00 6C 00 6C 00 6F
```

This is the Unicode (big-endian) encoding with no BOM. Notepad does not support this encoding.

```
FE FF 00 48 00 65 00 6C 00 6C 00 6F
```

This is the Unicode (big-endian) encoding with BOM. Notice that this BOM is in the opposite order from the little-endian BOM.

```
EF BB BF 48 65 6C 6C 6F
```

This is UTF-8 encoding. The first three bytes are the UTF-8 encoding of the BOM.

```
2B 2F 76 38 2D 48 65 6C 6C 6F
```

This is UTF-7 encoding. The first five bytes are the UTF-7 encoding of the BOM. Notepad doesn't support this encoding.

Notice that the UTF7 BOM encoding is just the ASCII string "+/v8-", which is difficult to distinguish from just a regular file that happens to begin with those five characters (as odd as they may be). The encodings that do not have special prefixes and which are still supported by Notepad are the traditional ANSI encoding (i.e., "plain ASCII") and the Unicode (little-endian) encoding with no BOM. When faced with a file that lacks a special prefix, Notepad is forced to guess which of those two encodings the file actually uses. The function that does this work is IsTextUnicode, which studies a chunk of bytes and does some statistical analysis to come up with a guess. And as the documentation notes, “Absolute certainty is not guaranteed.” Short strings are most likely to be misdetected.

[Raymond is currently on vacation; this message was pre-recorded.]

Raymond Chen

Follow