Subject: Re: LoadFile problem with accented chars
Posted by mirek on Mon, 09 Feb 2009 07:12:01 GMT
View Forum Message <> Reply to Message

koldo wrote on Sun, 08 February 2009 16:11Hello luzr

It seems it is a matter of Notepad itself. If the file has 7 bits chars there is no problem, but after

Using this test program:

```
CONSOLE_APP_MAIN
{
 String data = LoadFile("C:\\test.txt");
 for (int i = 0; i < data.GetCount(); ++i)
  puts(Format("%d: %d", i, data[i]));
 getchar();
}
```

```
0: 97
1: 45
2: -31
```

but after saving and opening the file some times, I get this:

```
0: -1
1: -2
2: 97
3: 0
4: 45
5: 0
6: -31
7: 0
```

and yesterday I got other output... The answer is that Notepad adds a "BOM" to the file if it thinks it requires a bigger encoding.

BOM (Byte Order Mark, http://unicode.org/faq/utf_bom.html#BOM) is a signature of letters in the begining of files that shows its encoding. For example:

- EF BB BF means UTF-8
- FF FE means UTF-16, little-endian

Why do not interpret it yourself?

I suggest implementing these:

```
WString LoadBOMW(const Stream& s);
WString LoadFileBOMW(const char *path);
void    SaveBOMUtf8(const Stream& s, const WString& data);
bool    SaveFileBOMUtf8(const char *path, const WString& data);

String  LoadBOM(const Stream& s); // Default encoding, usually utf-8
String  LoadFileBOM(const char *path);
void    SaveBOMUtf8(const Stream& s, const String& data);
bool    SaveFileBOMUtf8(const char *path, const String& data);
```

I would be glad to add them to Core.

Mirek
