
Subject: Unicode from file

Posted by [coolman](#) on Fri, 17 Feb 2023 18:54:01 GMT

[View Forum Message](#) <> [Reply to Message](#)

Hello,

I've been dealing with this for a while now and can't figure anything out. Could you please advise me?

I load the file using `FileIn in()` and `in.GetLine()` function. A string with escaped UTF "skryt\u00e9" is stored in the file. But when loading I get "skryt\\u00e9".
How should I convert input "skryt\\u00e9" to output "skryté".

Thanks in advance for your help

Subject: Re: Unicode from file

Posted by [coolman](#) on Sun, 19 Feb 2023 07:46:28 GMT

[View Forum Message](#) <> [Reply to Message](#)

Based on CParser, I created a simple functionality for decoding escape sequences for UTF. It's really simple, so in case UTF decoding fails, they don't translate this bad sequence, but leave it unchanged.

```
static bool ReadHex(StringStream &in, dword &hex, int n) {
    hex = 0;
    while (n--) {
        if (in.IsEof())
            return false;
        int c = in.Get();
        if (!IsXDigit(c))
            return false;
        hex = (hex << 4) + ctoi(c);
    }
    return true;
}
```

```
static String GetUtfSmall(StringStream &in) {
    String result;
    dword hex = 0;
    if (ReadHex(in, hex, 4)) {
        if (hex >= 0xD800 && hex < 0xDBFF) {
            int c = in.Get();
            int next = in.Get();
            if (c == '\\' && next == 'u') {
                dword hex2;
                if (ReadHex(in, hex2, 4) && hex2 >= 0xDC00 && hex2 <= 0xDFFF) {
```

```

        result.Cat(ToUtf8(((hex & 0x3ff) << 10) | (hex2 & 0x3ff) + 0x10000));
    }
}
} else {
    if (hex > 0 && hex < 0xDC00) {
        result.Cat(ToUtf8(hex));
    }
}
}
return result;
}

```

```

static String GetUtfCapital(StringStream &in) {
    String result;
    dword hex = 0;
    if (ReadHex(in, hex, 8) && hex > 0 && hex < 0x10ffff) {
        result.Cat(ToUtf8(hex));
    }
    return result;
}

```

```

static String DecodeEscapedUtf(const String &s) {
    StringStream ss(s);
    String result;

    while (!ss.IsEof()) {
        int c = ss.Get();
        if (c == '\\') {
            int next = ss.Get();
            int64 pos = ss.GetPos();
            String utf;
            switch (next) {
                case 'u':
                    utf = GetUtfSmall(ss);
                    break;
                case 'U':
                    utf = GetUtfCapital(ss);
                    break;
                default:
                    break;
            }
            if (utf.GetCount() > 0) {
                result.Cat(utf);
            } else {
                ss.Seek(pos);
                result.Cat(c);
                result.Cat(next);
            }
        }
    }
}

```

```
} else {  
    result.Cat(c);  
}  
}  
return result;  
}
```

Subject: Re: Unicode from file
Posted by [coolman](#) on Sun, 19 Feb 2023 14:39:01 GMT
[View Forum Message](#) <> [Reply to Message](#)

For clarification: The presented solution can be used to process strings that contain UTF escape sequences, when the resulting string will contain only translated characters.
