Subject: String w/high characters but not UTF?
Posted by jeremy_c on Sun, 08 Aug 2010 00:52:40 GMT

I am using a web service that returns data in a very old format (has endured since the DOS days!)... It's something like this:

FIELD1(char 181)FIELD2(char 182)FIELD3(char 184)FIELD4(char 185)

The (char 181) being actually chr(181). So, to parse you know you want the field starting with chr(183) ... for example ... You find 183, then search until 184 and take everything in between.

I seem to be having problems with U++ seeing that as a UTF string and doing weird things with it.

How can I prevent this? I am using HttpClient.Execute(); to get the content.

Jeremy

Subject: Re: String w/high characters but not UTF?
Posted by koldo on Tue, 10 Aug 2010 07:13:19 GMT

jeremy_c wrote on Sun, 08 August 2010 02:52I am using a web service that returns data in a very old format (has endured since the DOS days!)... It's something like this:

FIELD1(char 181)FIELD2(char 182)FIELD3(char 184)FIELD4(char 185)

The (char 181) being actually chr(181). So, to parse you know you want the field starting with chr(183) ... for example ... You find 183, then search until 184 and take everything in between.

I seem to be having problems with U++ seeing that as a UTF string and doing weird things with it.

How can I prevent this? I am using HttpClient.Execute(); to get the content.

Jeremy
Hello Jeremy

One question: The problem is when you get the data or when you parse the data?

Subject: Re: String w/high characters but not UTF?
Posted by jeremy_c on Tue, 10 Aug 2010 12:29:41 GMT

When I parse the data. I create a small U++ app that shows what my problem is:

```
#include <Core/Core.h>

using namespace Upp;

CONSOLE_APP_MAIN
{
  char data[] = { 65, 65, 65, 181, 65, 65, 65, 182, 65, 65, 183 };
  String d(data);
  for (int i=0; i < d.GetCount(); i++) {
    LOG(FormatInt(i) + "=" + FormatInt(d[i]));
  }
}
```

Thanks for any help with this. I'm sure it's simple but it's driving me nuts!

Jeremy

Subject: Re: String w/high characters but not UTF?
Posted by cbpporter on Tue, 10 Aug 2010 13:23:42 GMT

There is nothing wrong with that program, related to UTF8 or otherwise. I behaves as it should. The problem is that you are inserting a large value like 182 in a signed char and the result gets interpreted as a negative number.

Subject: Re: String w/high characters but not UTF?
Posted by koldo on Tue, 10 Aug 2010 14:01:38 GMT

cbpporter wrote on Tue, 10 August 2010 15:23There is nothing wrong with that program, related to UTF8 or otherwise. I behaves as it should. The problem is that you are inserting a large value like 182 in a signed char and the result gets interpreted as a negative number.

Yes.

For example compiling with MSC I got three warnings like this:

 warning C4309: 'initializing' : truncation of constant value

for the 181, 182 and 183.

In addition String d does not know the length of char data[] as it is not ended with '\0'. This easily can produce an error.

Check this:

```
#include <Core/Core.h>

using namespace Upp;

CONSOLE_APP_MAIN
{
 {
  puts("Original");
  char data[] = { 65, 65, 65, 181, 65, 65, 65, 182, 65, 65, 183 };
  String d(data);
  for (int i=0; i < d.GetCount(); i++)
   puts(FormatInt(i) + "=" + FormatInt(d[i]));
 }
 {
  puts("Changed");
  byte data[] = { 65, 65, 65, 181, 65, 65, 65, 182, 65, 65, 183 };
  String d(data, 11);
  for (int i=0; i < d.GetCount(); i++)
   puts(FormatInt(i) + "=" + FormatInt(byte(d[i])));
 }
 getchar();
}
```

The output is this:

Original
0=65
1=65
2=65
3=-75
4=65
5=65
6=65
7=-74
8=65
9=65
10=-73
Changed
0=65
1=65
2=65

3=181
4=65
5=65
6=65
7=182
8=65
9=65
10=183

byte type is a natural way in U++ to handle binary data.

If you need a classic C array with undefined length in compiling time you can also use:

Buffer<byte> data;

data.Alloc(dataLen);

instead of the usual and more dangerous malloc/free/new/delete.

Subject: Re: String w/high characters but not UTF?
Posted by jeremy_c on Tue, 10 Aug 2010 14:21:07 GMT
View Forum Message <> Reply to Message

The problem is that I am using HttpClient to get this data. So it actually looks like:

String data = HttpClient(...);

data then includes the positive and negative characters.

The original question was how to get this data correctly or to deal with it once it has been gotten incorrectly.

Jeremy

Subject: Re: String w/high characters but not UTF?
Posted by cbpporter on Tue, 10 Aug 2010 14:29:17 GMT
View Forum Message <> Reply to Message

Well you could always use two's complement. Sorry, I have to go right now but I'll try and explain latter.

Subject: Re: String w/high characters but not UTF?
Posted by jeremy_c on Tue, 10 Aug 2010 19:15:28 GMT

I got it figured out. I just cast it to an unsigned char.

Jeremy

Subject: Re: String w/high characters but not UTF?
Posted by koldo on Wed, 11 Aug 2010 06:28:24 GMT

jeremy_c wrote on Tue, 10 August 2010 21:15I got it figured out. I just cast it to an unsigned char.

Jeremy

Remember this line :

```
puts(FormatInt(i) + "=" + FormatInt(byte(d[i])));
```