2021_11_16

November 17, 2021

1 2021-11-16

1.0.1 Corso ITS

1.1 Magento & e-commerce software

1.2 ### Fondamenti di Programmazione (Andrea Ribuoli)

[1]:
```
print('\u0043\u0044');
```

```
CD
```

[2]:
```
print('\u20AC');
```

```
€
```

[3]:
```
print('\U00010001');
```

[4]:
```
print('\U00105252');
```

[5]:
```
print('\U0010FFFF');
```

[6]:
```
print('\U00110000');
```

```
  File "<ipython-input-6-fcf8aa67a823>", line 1
    print('\U00110000');
          ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes in␣
 ↪position 0-9: illegal Unicode character
```

[21]:
```
!hexdump -h
```

```
Usage:
```

```
 hexdump [options] <file>…

Display file contents in hexadecimal, decimal, octal, or ascii.

Options:
 -b, --one-byte-octal      one-byte octal display
 -c, --one-byte-char       one-byte character display
 -C, --canonical           canonical hex+ASCII display
 -d, --two-bytes-decimal   two-byte decimal display
 -o, --two-bytes-octal     two-byte octal display
 -x, --two-bytes-hex       two-byte hexadecimal display
 -L, --color[=<mode>]      interpret color formatting specifiers
                             colors are enabled by default
 -e, --format <format>     format string to be used for displaying data
 -f, --format-file <file>  file that contains format strings
 -n, --length <length>     interpret only length bytes of input
 -s, --skip <offset>       skip offset bytes from the beginning
 -v, --no-squeezing        output identical lines

 -h, --help                display this help
 -V, --version             display version

For more details see hexdump(1).
```

Abbiamo determinato l'encoding UTF-8 del simbolo € a partire dal fatto che ad esso è stato assegnato il **code point** `U+20AC` in Unicode.

- per prima cosa abbiamo identificato il range di appartenenza: `U+0800 - U+FFFF` che ha determinato l'uso di **3 byte**
- nota allora la struttura in bit `1110xxxx-10xxxxxx-10xxxxxx` abbiamo valorizzato i 16 bit variabili sulla base dei 4 semi-byte del code point `20AC`
- indichiamo `1110xxxx-10xxxxxx-10xxxxxx` come `1110xxxx-10yyyyww-10wwzzzz` essendo rispettivamente xxxx la notazione binaria dell'esadecimale **2**, yyyy quella dell'esadecimale **0**, wwww di **A** e zzzz di **C**
- poiché xxxx = 0010, yyyy = 0000, wwww = 1010 e zzzz = 1100 otteniamo `1110xxxx-10yyyyww-10wwzzzz` = `1110010-10000010-10101100`
- ora `1110010-10000010-10101100` corrisponde a `E2-82-AC` in notazione esadecimale

Abbiamo verificato l'impatto sulla dimensione di un file sorgente a seguito della sostituzione di una lettera **e** con il carattere **€** usando il comando Unix `ls -la`.

Abbiamo identificato la sequenza `E2-82-AC` a fronte di un `hexdump` del file sorgente in oggetto.

Abbiamo condotto uno studio approfondito della funzione main simulando il comportamento del compilatore nel preparare la struttura che indichiamo con `char *argv[]` nel template standard della funzione *main()* stessa.