Zeichendarstellung mit Unicode und UTF-8

ASCII – American Standard Code for Information Interchange

In der Mittelstufe wurde die Codierung von Text mithilfe des ASCII-Standards besprochen. Hierbei wird jedem Zeichen ein Wert zwischen 0 und 255 (8 Bit) zugewiesen. Oben siehst du die ASCII-Codetabelle, leere Zellen enthalten Steuerzeichen, welche für die Darstellung am PC nötig waren. Die wichtigsten Steuerzeichen sind in der Tabelle beschrieben.

In einem früheren, hauptsächlich in Amerika benutzten Standard waren lediglich die Zeichen von 0 bis 127 definiert, das letzte, achte Bit wurde zur Fehlerüberprüfung verwendet. Erst später wurde das 8. Bit dazu genommen, um weitere Zeichen, wie z.B. die deutschen Umlaute codieren zu können.

(A1)

Wandle die nachfolgenden Wörter, die in Hexadezimal-Darstellung vorliegen, in lesbaren Text um:

1. `49 6E 66 6F 72 6D 61 74 69 6B`
2. `42 69 6E E4 72`
3. `43 6F 6D 70 75 74 65 72`

Mit einer 8-Bit-Codierung lassen sich nicht mehr Zeichen darstellen, was insbesondere bei anderen Sprachen – wie z.B. griechisch – andere Codierungen nötig machte. Da in diesen Sprachen jedoch die bei uns gebräuchlichen Umlaute nicht benötigt werden, wurde der durch das 8. Bit hinzugekommene Block vom Zeichen 128 bis 255 für die dortigen Zeichen verwendet. Diese und andere länderspezifischen Codierungen lassen sich z.B. unter https://de.wikipedia.org/wiki/ISO_8859 nachschauen.

(A2)

Welche der obigen Wörter würden mit den griechischen Zeichensatz falsch dargestellt werden und warum?

Unicode

Um Probleme, die sich zum einen mit unterschiedlichen Zeichensätzen, zum anderen auch durch andere Sprachen, die mehr als 128 Zeichen umfassen, ergeben haben, wurde der Unicode-Standard entwickelt.

> Unicode ist ein internationaler Standard, in dem langfristig für jedes sinnvolle Schriftzeichen oder Textelement aller bekannten Schriftkulturen und Zeichensysteme ein digitaler Code festgelegt wird. Ziel ist es, die Verwendung unterschiedlicher und inkompatibler Kodierungen in verschiedenen Ländern oder Kulturkreisen zu beseitigen. Unicode wird ständig um Zeichen weiterer Schriftsysteme durch das Unicode-Konsortium ergänzt. (Wikipedia, https://de.wikipedia.org/wiki/Unicode)

Im Unicode Standard hat jedes Zeichen einen eigenen „Unicode-Code“, damit lassen sich derzeit 1.111.998 elementare Zeichen („Codepunkte“) abbilden. Darstellung: U+00DF (Mindestens 4x4Bit, bis zu U+10FFFF) Diese Codepunkte bilden den Unicode Zeichensatz.

Die Zeichen des Zeichensatzes werden wiederum auf unterschiedliche Weisen codiert, beispielsweise in Betriebssystemen. Wir betrachten die UTF-8 Kodierung von Unicode Zeichen genauer.

UTF-8 Implementation des Unicode Zeichensatzes

Hier kann ein einzelnes Zeichen in der UTF-8-Codierung bis zu 4 Bytes umfassen, nach folgenden Regeln:

- Ist die Binärdarstellung des Unicode-Codes nicht länger als ein Byte und das das erste Bit eine 0, werden die restlichen 7 Bit gemäß des ASCII Codes verwendet, die 128 verbleibenden Möglichkeiten entsprechen also genau dem ASCII-Code.
- Ist die Binärdarstellung des Unicode-Codes länger als ein Byte oder der Code ist ein Byte lang und beginnt mit einer 1 geht man wie folgt vor: Der Unicode-Code wird in 6 Bit lange Teile aufgeteilt. Für jedes dieser 6 Bit Pakete wird ein Byte zur Darstellung verwendet, jedes Byte beginnt mit '10'. Das erste Byte beginnt mit einer '1' für jedes Byte, das verwendet wird. Benötigt man also 3 Byte, um ein Zeichen in UTF-8 darzustellen, beginnt das erste Byte mit '111'. Bevor die Nutzdaten beginnen, muss noch eine Null eingefügt werden [1)]

Beispiele:

(1)

$y = 79_{16} = 0111\ 100_2$

Beginnt mit einer Null und ist nicht länger als ein Byte → die letzten 7Bit werden verwendet, um zu codieren, also ein „ASCII k“ in UTF-8

UTF-8: 0110 1011

(2)

ä = $E4_{16}$ = $1110\ 0100_2$

Nur ein Byte lang, beginnt aber mit einer 1. Die 8 Bit müssen in 6 Bit Abschnitte geteilt und auf 2 Byte verteilt werden, beginnend von rechts, links wird stets mit 0en aufgefüllt:

`000011 100100`

- Das zweite Byte beginnt nach den Regeln mit 10, daran schließen die Nutzdaten an: `10 100100`
- Das erste Byte beginnt mit 11 (weil man zwei Byte benötigt) dann wird mit 0en aufgefüllt, dann kommen die Nutzedaten: `11 000011`

Die UTF-8 Codierung des Unicode-ä ist also `1100 0011 1010 0100`. Die Nutzdaten, die den Code des Unicode Zeichens transportieren sind in jeden Byte nur die jeweils letzten 6 Bit.

(3) 乐 → U+4E50 → $4E50_{16}$ → $0100\ 1110\ 0101\ 0000_2$

- 16 Bit Daten zu codieren, dafür braucht man 3 Byte (3 x 6 = 18)
- Der UTF-8 Code beginnt also mit der Startsequenz 111
- Dann von rechst beginnend 6 Bit (01 000), das Byte beginnt mit 10 (Regel) also ist das dritte Byte `1010 1000`
- Die nächsten 6 Bit analog: `1110 01` → `1011 1001`
- Die fehlenden 4 Bit `0100` mit Padding + Startsequenz (111)ergeben das erste Byte `1110 0100`

Die UTF-8 Codierung des Unicode-Zeichens 乔 ist also 3 Byte lang und sieht so aus: `1110 0100 1011 1001 1010 0000`

(A3)

Wandle die nachfolgenden Zeichen des Unicode Zeichensatzen in die UTF-8-Codierung um. Der Hexadezimalcode des Unicode Zeichens ist jeweils angegeben.

Gehe jeweils wie in den Beispielen oben vor. Markiere die „Nutzdaten“ die das eigentlich Unicode-Zeichen „transportieren“.

1. I=49_{16}
2. Ö=$D6_{16}$
3. 弈=$5F08_{16}$
4. □=$1F60A_{16}$

Lösung 1

0**1001001**, ein Byte, erstes Bit 0.

Lösung 2

110**00011** 10**010110**

Lösung 3

1110**0101** 10**111100** 10**001000**

Lösung 4

1111**0000** 10**011111** 10**011000** 10**001010**

(A4)

Wie viele unterschiedliche Unicode-Zeichen lassen sich theoretisch mit 1 Byte, 2 Bytes, 3 Bytes und 4 Bytes unter Beachtung der UTF-8-Regeln darstellen?

Lösung

- 1 Byte: 7 nutzbare Bits → 2^7 = 128 Zeichen
- 2 Bytes: 5+6 = 11 nutzbare Bits → 2^{11} = 2 048 Zeichen
- 3 Bytes: 4+6+6 = 16 nutzbare Bits → 2^{16} = 65 536 Zeichen
- 4 Bytes: 3+6+6+6 = 21 nutzbare Bits → 2^{21} = 2 097 152 Zeichen

---

1)
Warum?