Παραδείγματα κωδικοποίησης Unicode UTF-8

Κωδικοποίηση και αποκωδικοποίηση
Το πρότυπο Unicode αναθέτει έναν αριθμό (**codepoint**) σε κάθε χαρακτήρα των αλφαβήτων που υποστηρίζει. Για παράδειγμα, ο λατινικός χαρακτήρας **L** αντιστοιχεί στον δεκαεξαδικό αριθμό **4C**, ο ελληνικός χαρακτήρας **ψ** στον δεκαεξαδικό αριθμό **3C8** κ.ο.κ.

Κατά την εκτέλεση ενός προγράμματος που χειρίζεται κείμενο Unicode, οι αριθμοί αυτοί βρίσκονται στη μνήμη του υπολογιστή, καταλαμβάνοντας 2 ή 4 bytes ο καθένας. Κατά την αποθήκευση όμως σε αρχείο χρησιμοποιείται η κωδικοποίηση UTF-8 που μετατρέπει τους αριθμούς Unicode σε μια σειρά από bytes με καθορισμένη μορφή και σειρά.

Η μετατροπή από αριθμό Unicode σε bytes σύμφωνα με το πρότυπο UTF-8 ονομάζεται **κωδικοποίηση**, ενώ η μετατροπή από τα bytes ενός αρχείου σε αριθμούς Unicode ονομάζεται **αποκωδικοποίηση**.

Το πρότυπο UTF-8

Το πρότυπο UTF-8 προβλέπει τη μετατροπή αριθμών Unicode σε σειρές από bytes **μεταβλητού μήκους**, ανάλογα με τον κάθε αριθμό, σύμφωνα με τον επόμενο πίνακα:

Περιοχή αριθμού Unicode	**Κωδικοποίηση UTF-8**
0...7F	0xxxxxxx
80...7FF	110xxxxx 10xxxxxx
800...FFFF	1110xxxx 10xxxxxx 10xxxxxx
10000...10FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

Παρατηρήσεις

1. Οι χαρακτήρες του παλιού κώδικα ASCII (1η γραμμή πίνακα) παραμένουν ως έχουν. Αυτό σημαίνει πως κάθε αρχείο ASCII είναι και έγκυρο αρχείο UTF-8.
2. Οι κωδικοποιήσεις (δεξιά στήλη πίνακα) ξεκινούν με διαφορετικά bits ανά σειρά.
3. Όταν υπάρχουν πολλαπλά bytes στην κωδικοποίηση (δεξιά στήλη πίνακα), το πρώτο byte ακολουθείται από έναν αριθμό συνοδευτικών bytes, τα οποία ξεκινούν με τον συνδυασμό bits 10.…

Παράδειγμα κωδικοποίησης αριθμών Unicode σε σειρά bytes

Κωδικοποιήστε κατά UTF-8 τους **δεκαεξαδικούς** αριθμούς α) **32** (αντιστοιχεί στο ψηφίο 2) και β) **386** (ο ελληνικός χαρακτήρας Ά).

Δεκαεξαδικοί αριθμοί
Τα παραδείγματα δίνονται στο **δεκαεξαδικό σύστημα** για την εύκολη μετατροπή σε σειρές από bits. Θυμηθείτε ότι κάθε δεκαεξαδικό ψηφίο είναι 4 bits:

```
0 → 0000  1 → 0001  2 → 0010  3 → 0011
4 → 0100  5 → 0101  6 → 0110  7 → 0111
8 → 1000  9 → 1001  A → 1010  B → 1011
C → 1100  D → 1101  E → 1110  F → 1111
```

Εξετάζουμε **την αριστερή στήλη** (“Περιοχή αριθμού Unicode”) του πίνακα κωδικοποίησης UTF-8 και, ανάλογα με την περιοχή που βρίσκεται ο αριθμός που μελετάμε, επιλέγουμε την αντίστοιχη κωδικοποίηση **στη δεξιά στήλη** (“Κωδικοποίηση UTF-8”):

α) Ο δεκαεξαδικός αριθμός **32** βρίσκεται στην περιοχή **0...7F** (00 < 32 < 7F), συνεπώς ακολουθούμε την κωδικοποίησης της **πρώτης γραμμής** του πίνακα:

32 (δεκαεξαδικό) = 00110010

Τα 7 χαμηλότερα bits (τα σκιασμένα) μεταφέρονται στις αντίστοιχες θέσεις (σημειωμένες ως x) στο δεξιό μέρος του πίνακα 0xxxxxxx και προκύπτει μετά την αντικατάσταση ο αριθμός 00110010, δηλαδή το δεκαεξαδικό **32**.

Σημ: ειδικά για την πρώτη γραμμή του πίνακα ο αριθμός μένει ως έχει κατά την κωδικοποίηση. Όπως είπαμε στην αρχή, πρόκειται για την περιοχή των χαρακτήρων ASCII που μένει αναλλοίωτη!

β) Ο δεκαεξαδικός αριθμός **386** βρίσκεται στην περιοχή **80...7FF** (080 < 386 < 7FF), άρα χρησιμοποιούμε τη **δεύτερη γραμμή** του πίνακα για την κωδικοποίηση και προκύπτουν 2 bytes εξόδου.

Εδώ θα τοποθετήσουμε τα 11 χαμηλότερα bits (φαίνονται σκιασμένα στο επόμενο) στις αντίστοιχες θέσεις της κωδικοποίησης:

386 (δεκαεξαδικό) = 00111000110

Κωδικοποίηση = 110xxxxx 10xxxxxx και, μετά τη συμπλήρωση, 11001110 10000110 ή δεκαεξαδικά τα bytes **CE 86**.

Παράδειγμα αποκωδικοποίησης σειράς bytes σε αριθμούς Unicode

Έστω ότι από ένα αρχείο κειμένου κατά UTF-8 διαβάζουμε τα εξής bytes:

`66 CF 85 E5 AD 9A` (δεκαεξαδικό)

Πόσοι χαρακτήρες Unicode είναι; Ποιος ο αριθμός Unicode του κάθε χαρακτήρα;

Εξετάζουμε τη μορφή των bytes στη δεξιά στήλη (“Κωδικοποίηση UTF-8”) του πίνακα κωδικοποίησης UTF-8:

1. Διαβάζουμε το πρώτο byte του αρχείου, αυτό είναι το δεκαεξαδικό **66** ή δυαδικά 01100110.

Από το περισσότερο σημαντικο bit 0 καταλαβαίνουμε ότι βρισκόμαστε στην πρώτη γραμμή του πίνακα (αντιστοιχεί στο 0xxxxxxx). Συνεπώς ο αριθμός Unicode **του πρώτου χαρακτήρα** είναι το byte **66** ως έχει (ο χαρακτήρας **f**).

2. Διαβάζουμε το επόμενο byte, το δεκαεξαδικό **CF** ή δυαδικά 11001111.

Από τα περισσότερο σημαντικά bits 110 καταλαβαίνουμε ότι είμαστε στη δεύτερη γραμμή του πίνακα (αντιστοιχεί στο 110xxxxx) και ότι χρειαζόμαστε **ακόμα ένα byte**, το δεκαεξαδικό **85** που ακολουθεί στο αρχείο.

Συνολικά, τα δύο bytes στο δυαδικό είναι τα 11001111 10000101 και αν αφαιρέσουμε τα bits της κωδικοποίησης (σκιασμένα bits) μένει η καθαρή πληροφορία του αριθμού Unicode **του δεύτερου χαρακτήρα**:

01111000101 ή στο δεκαεξαδικό **3C5** (ο χαρακτήρας **υ**).

3. Διαβάζουμε το επόμενο byte, το δεκαεξαδικό **E5** ή δυαδικά 11100101.

Από τα περισσότερο σημαντικά bits 1110 καταλαβαίνουμε ότι είμαστε στην τρίτη γραμμή του πίνακα (αντιστοιχεί στο 1110xxxx) και ότι χρειαζόμαστε **ακόμα δύο bytes**, τα δεκαεξαδικά **AD** και **9A** που ακολουθούν.

Τα τρία bytes στο δυαδικό είναι τα 11100101 10101101 10011010 και αν αφαιρέσουμε τα bits της κωδικοποίησης (σκιασμένα bits) μένει η καθαρή πληροφορία του αριθμού Unicode **του τρίτου χαρακτήρα**:

0101101101011010 ή στο δεκαεξαδικό **5B5A** (ο χαρακτήρας **孚**).