

Original by Markus Kuhn, adapted for HTML by Martin Dürst.

UTF-8 encoded sample plain-text file

Markus Kuhn ['maʊ̯kəs ku:n] <mkuhn@acm.org> – 1999-08-20

The ASCII compatible UTF-8 encoding of ISO 10646 and Unicode plain-text files is defined in RFC 2279 and in ISO 10646-1 Annex R.

Using Unicode/UTF-8, you can write in emails and source code things such as

Mathematics and Sciences:

$$\square E \cdot da = Q, \quad n \rightarrow \infty, \quad \sum f(i) = \prod g(i), \quad \forall x \in \mathbb{R}: \lceil x \rceil = -\lfloor -x \rfloor, \quad \alpha \wedge \neg \beta = \neg(\neg \alpha \vee \beta),$$

$$\mathbb{N} \subseteq \mathbb{N}_0 \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}, \quad \perp < a \neq b \equiv c \leq d \quad \square T \Rightarrow (A \Leftrightarrow B),$$

$2\text{H}_2 + \text{O}_2 \rightleftharpoons 2\text{H}_2\text{O}$, R = 4.7 kΩ, Ø 200 mm

Linguistics and dictionaries:

ði ɪntə'næʃənəl fə'nɛtik əsoussi'eɪʃn
Y ['Ypsilɔn], Yen [jɛn], Yoga ['jo:gə]

APL:

$$((\nabla \iota V) = \iota \rho V) / V \leftarrow, V \quad \square \leftarrow \iota \rightarrow \rho \Delta \nabla \supset \neg \square \square$$

Nicer typography in plain text files:

- ‘single’ and “double” quotes

- Curly apostrophes: “We’ve been here”
- Latin-1 apostrophe and accents: ‘ ’`
- „deutsche“ „Anführungszeichen“
- †, ‡, ‰, •, 3–4, –, –5/+5, ™, ...
- ASCII safety test: 1lI|, 00D, 8B
- the euro symbol: 14.95 €

Greek (in Polytonic):

The Greek anthem:

Σὲ γνωρίζω ἀπὸ τὴν κόψη
τοῦ σπαθιοῦ τὴν τρομερή,
σὲ γνωρίζω ἀπὸ τὴν ὅψη
ποὺ μὲ βία μετράει τὴ γῆ.

’Απ’ τὰ κόκκαλα βγαλμένη
τῶν ’Ελλήνων τὰ ιερά
καὶ σὰν πρῶτα ἀνδρειωμένη
χαῖρε, ὦ χαῖρε, ’Ελευθεριά!

From a speech of Demosthenes in the 4th century BC:

Οὐχὶ ταῦτὰ παρίσταται μοι γιγνώσκειν, ὡς ἄνδρες ’Αθηναῖοι,
ὅταν τ’ εἰς τὰ πράγματα ἀποβλέψω καὶ ὅταν πρὸς τοὺς
λόγους οὖς ἀκούω· τοὺς μὲν γὰρ λόγους περὶ τοῦ
τιμωρήσασθαι φίλιππον ὄρῳ γιγνομένους, τὰ δὲ πράγματ’
εἰς τοῦτο προήκοντα, ὃσθ’ ὅπως μὴ πεισόμεθ’ αὐτοὶ
πρότερον κακῶς σκέψασθαι δέον. οὐδέν οὖν ἄλλο μοι δοκοῦσιν
οἱ τὰ τοιαῦτα λέγοντες ἢ τὴν ὑπόθεσιν, περὶ ᾧς βουλεύεσθαι,
οὐχὶ τὴν οὗσαν παριστάντες ὑμῖν ἀμαρτάνειν. ἐγὼ δέ, ὅτι μέν

ποτ' ἔξῆν τῇ πόλει καὶ τὰ αὐτῆς ἔχειν ἀσφαλῶς καὶ φίλιππον τιμωρήσασθαι, καὶ μάλ' ἀκριβῶς οἶδα· ἐπ' ἔμοῦ γάρ, οὐ πάλαι γέγονεν ταῦτ' ἀμφότερα· νῦν μέντοι πέπεισμαι τοῦθ' ἰκανὸν προλαβεῖν ἡμῖν εἶναι τὴν πρώτην, ὅπως τοὺς συμμάχους σώσομεν. ἐὰν γὰρ τοῦτο βεβαίως ὑπάρξῃ, τότε καὶ περὶ τοῦ τίνα τιμωρήσεται τις καὶ ὃν τρόπον ἔξεσται σκοπεῖν· πρὶν δὲ τὴν ἀρχὴν ὁρθῶς ὑποθέσθαι, μάταιον ἥγοῦμαι περὶ τῆς τελευτῆς ὄντινοῦν ποιεῖσθαι λόγον.

Δημοσθένους, Γ' Ὀλυνθιακὸς

Georgian:

From a Unicode conference invitation:

გთხოვთ ახდავე გაიაროთ რეგისტრაცია Unicode-ის მეათე საერთაშორისო კონფერენციაზე დასასწრებად, რომელიც გაიმართება 10-12 მარტს, ქ. მაინცი, გერმანიაში. კონფერენცია შეჰვრებს ერთად მსოფლიოს ექსპერტებს ისეთ დარგებში როგორიცაა ინტერნეტი და Unicode-ი, ინტერნაციონალიზაცია და ღოვანიზაცია, Unicode-ის გამოყენება თვერაციულ სისტემებსა, და გამოყენებით პროგრამებში, შრიფტებში, ტექსტების დამუშავებასა და მრავალენოვან კომპიუტერულ სისტემებში.

Russian:

From a Unicode conference invitation:

Зарегистрируйтесь сейчас на Десятую Международную Конференцию по Unicode, которая состоится 10-12 марта 1997 года в Майнце в Германии. Конференция соберет широкий круг экспертов по вопросам глобального Интернета и Unicode, локализации и интернационализации, воплощению и применению Unicode в различных операционных системах и программных приложениях, шрифтах, верстке и многоязычных компьютерных системах.

Thai (UCS Level 2):

Excerpt from a poetry on The Romance of The Three Kingdoms (a Chinese classic 'San Gua'):

The diagram consists of two groups of horizontal bars. The left group contains 10 bars, each composed of 10 small squares. The right group contains 11 bars, also each composed of 10 small squares. A single vertical line is positioned between the two groups, creating a central vertical boundary.

(The above is a two-column text. If combining characters are handled correctly, the lines of the second column should be aligned with the | character above.)

Ethiopian:

Proverbs in the Amharic language:

The image shows a grid of 40 sets of five empty boxes each. The boxes are arranged in 8 rows, with each row containing 5 sets of boxes. Each set consists of five adjacent horizontal rectangles. The entire grid is composed of black outlines on a white background.

Runes:

The image shows a horizontal row of 12 different rectangles. Each rectangle is divided into 4 equal-sized smaller rectangles. The ways in which the main rectangle is divided vary across the row, illustrating different fraction models for the same fraction.

(Old English, which transcribed into Latin reads 'He cwaeth that he bude thaem lande northweardum with tha Westsae.' and means 'He said that he lived in the northern land near the Western Sea.')

Braille:

The image shows four separate groups of four empty rectangular boxes. Each group is intended for a single digit, such as tens or ones. The boxes are arranged horizontally in two rows of two.

A 6x10 grid of 60 empty rectangular boxes arranged in six rows and ten columns. The boxes are outlined in black and have a thin white interior.

The image shows a large grid of 100 small squares arranged in a 10x10 pattern. Each square is a 10x10 grid itself, creating a total of 1000 individual 10x10 boxes. This visual representation is commonly used in digital image processing to show a convolutional neural network's receptive field or a feature map.

(The first couple of paragraphs of "A Christmas Carol" by Dickens)

Compact font selection example text:

ABCDEFGHIJKLMNOPQRSTUVWXYZ /0123456789
abcdefghijklmnopqrstuvwxyz €©µÀÆÖÞÞéöý
— „” „†•‰™œšÝž€ АВГДΩαβγδω АБВГДабвгд

forall i < len
 if arr[i] == target:
 return i
return -1

Greetings in various languages:

Hello world, Καλημέρα κόσμε, こんにちは

Box drawing alignment tests:

