EBook Gratuito

APPRENDIMENTO unicode

Free unaffiliated eBook created from
Stack Overflow contributors.

#unicode

Sommario

Di

Capitolo 1: Iniziare con Unicode

Osservazioni

Lo standard Unicode è un set di caratteri standardizzato internazionale. Tenta di assegnare caratteri e simboli da ogni sistema di scrittura a un numero univoco. Con tutte le principali nuove versioni, ulteriori caratteri vengono aggiunti allo standard per raggiungere questo obiettivo. Fornendo un set di caratteri unificato per tutti i sistemi di scrittura, le informazioni di testo possono essere scambiate in un formato Unicode indipendente da qualsiasi piattaforma.

Lo standard Unicode contiene anche dati di proprietà sui caratteri e definisce algoritmi su come manipolare correttamente i caratteri. Ad esempio, questi algoritmi forniscono il metodo corretto per cercare e visualizzare il testo Unicode.

Versioni

Versione	Data di rilascio
2.0.0	1996/07/01
3.0.0	1999/09/01
3.1.0	2001/03/01
3.2.0	2002/03/01
4.0.0	2003-04-01
4.0.1	2004-03-01
4.1.0	2005-03-31
5.0.0	2006-07-14
5.1.0	2008-04-04
5.2.0	2009-10-01
6.0.0	2010-10-11
6.1.0	2012-01-31
6.2.0	2012/09/26
6.3.0	2013/09/30
7.0.0	2014/06/16

Versione	Data di rilascio
8.0.0	2015/06/17
9.0.0	2016/06/21

Examples

Installazione o configurazione

Istruzioni dettagliate su come installare o installare unicode.

Leggi Iniziare con Unicode online: https://riptutorial.com/it/unicode/topic/3188/iniziare-con-unicode

Capitolo 2: I caratteri possono essere costituiti da più punti di codice

Osservazioni

Un punto di codice Unicode, quello che i programmatori spesso pensano di un personaggio, spesso corrisponde a quello che l'utente pensa sia un personaggio. A volte tuttavia un "carattere" è costituito da più punti di codice, come mostrano gli esempi sopra.

Ciò significa che operazioni come tagliare una stringa o ottenere un carattere in un determinato indice potrebbero non funzionare come previsto. Ad esempio il 4 ° carattere della stringa `"Café "` è `'e'` (senza l'accento). Allo stesso modo, tagliando la corda alla lunghezza 4 si rimuoverà l'accento.

Il termine tecnico per un tale gruppo di punti di codice è un *grafo grapheme* . Vedi UAX # 29: Segmentazione del testo Unicode

Examples

segni diacritici

Una lettera con un segno diacritico può essere rappresentata con la lettera e una lettera di modifica combinatoria. Normalmente pensi ad `é` come a un personaggio, ma in realtà sono 2 punti di codice:

- `U+0065` - LETINA PICCOLA LETTERA E
- `U+0301` - COMBINAZIONE ACCENTO ACUTA

Allo stesso modo `ç` = `c` + `¸` , e `å` = `a` + `°`

forme combinate

Per complicare le cose, c'è spesso anche un punto di codice per la forma composta:

```
"Café " = 'C' + 'a' + 'f' + 'e' + '´'
"Café" = 'C' + 'a' + 'f' + 'é'
```

Sebbene queste stringhe siano uguali, non sono uguali e non hanno nemmeno la stessa lunghezza (5 e 4 rispettivamente).

Zalgo Text

C'è questa cosa chiamata Zalgo Text che lo spinge all'estremo. Ecco il primo grafo grapheme dell'esempio. Consiste di 15 punti di codice: la lettera latina `H` e 14 segni combinati.

h

Sebbene ciò non compaia nel testo normale, mostra che un "carattere" può realmente consistere in un numero arbitrario di punti di codice

Emoji e bandiere

Un sacco di emoji consistono in più di un punto di codice.

- : Un flag è definito come una coppia di "lettere indicatrici di simboli regionali" (+)
- : Alcune emoji possono essere seguite da un modificatore del tono della pelle: +
- o : Windows 10 consente di specificare se un'emoji è colorata o nero / bianco aggiungendo un selettore di variazione (`U+FE0E` o `U+FE0F`)
- : una famiglia. È stato codificato unendo le emoji a ragazzo, ragazza, donna e uomo (, , ,) insieme a joiners a larghezza zero (`U+200D`). Sulle piattaforme che lo supportano, questo è reso come un'emoji di una famiglia con due bambini.

Leggi I caratteri possono essere costituiti da più punti di codice online: https://riptutorial.com/it/unicode/topic/6485/i-caratteri-possono-essere-costituiti-da-piu-punti-di-codice

Capitolo 3: Il testo inglese non è solo ASCII

Osservazioni

Un'ipotesi che appare regolarmente è che quando si ha a che fare solo con il testo in inglese, è improbabile che incontri caratteri al di fuori del set di caratteri ASCII. Per evitare problemi con la gestione di Unicode correttamente, le persone sono tentate di fare cose come la rimozione di caratteri non ASCII o la rimozione di qualsiasi accento sulle lettere.

Questi esempi mostrano che questa ipotesi è sbagliata, e anche per il testo inglese dovresti fare attenzione a gestire correttamente i caratteri Unicode.

Examples

segni diacritici

Il testo inglese ha i segni diacritici occasionali.

- Parole di prestito, come née, café, entrée
- Nomi come Noël e Chloë
- Metti nomi, come Montréal e Québec

emoji

Emoji è abbastanza popolare con i social media in questi giorni.

- : `U+2603` - PUPAZZO DI NEVE
- : `U+01F600` - VISO DI GRINVIO
- : `U+01F42A` - CAMMELLO DROMEDARIO

Nota che la maggior parte delle emoji è al di fuori del piano multilingue di base. Molte nuove aggiunte consistono in più di un punto di codice:

- : Una bandiera è definita come una coppia di "lettere indicatrici di simboli regionali"
- : Questa è un'emoji più un modificatore della tonalità della pelle: +
- o : Windows 10 consente di specificare se un'emoji è colorata o nero / bianco aggiungendo un selettore di variazione (`U+FE0E` o `U+FE0F`)

Punteggiatura

Quasi tutti i testi scritti presentano segni di punteggiatura al di fuori del set di caratteri ASCII:

- trattini: il trattino - e il trattino -
- Virgolette: "virgolette" anziché "virgolette"
- I puntini di sospensione ...

Simboli speciali

Ci sono alcuni simboli comuni in uso:

- segno d'autore © e segni di marchio ® ™
- frazioni come ¼
- apici. Ad esempio, una scorciatoia per metri quadrati è m².

Leggi Il testo inglese non è solo ASCII online: https://riptutorial.com/it/unicode/topic/5198/il-testo-inglese-non-e-solo-ascii

Capitolo 4: UTF-8 come un modo di codifica di Unicode

Osservazioni

Che cos'è UTF-8 ?

UTF-8 è una codifica, che ha una lunghezza variabile e utilizza unità di codice a 8 bit: ecco perché UTF- **8** . In Internet UTF-8 è la codifica dominante (prima del 2008 ASCII era, ehich può anche gestire qualsiasi punto di codice Unicode.).

UTF-8 è uguale a Unicode?

"Unicode" non è una codifica - è un set di caratteri codificati - cioè un set di caratteri e una mappatura tra i caratteri e i punti di codice intero che li rappresentano. Ma molta documentazione lo usa per riferirsi alle *codifiche* . Su Windows, ad esempio, il termine Unicode viene utilizzato per fare riferimento a UTF-16.

UTF-8 è solo uno dei modi per codificare Unicode e come codifica converte le sequenze di byte in sequenze di caratteri e viceversa. UTF-16 e -32 sono altri formati di trasformazione Unicode.

BOM di UTF-8

Tutti e tre possono avere uno specifico Byte Order Mark, che essendo un numero magico segnala diverse cose importanti a un programma (ad esempio, Notepad ++) - ad esempio, il fatto che il flusso di testo importato è Unicode; inoltre aiuta a rilevare l'arte di Unicode usato per questo stream. Tuttavia, il consorzio Unicode consiglia di memorizzare UTF-8 senza firma. Alcuni software, ad esempio il compilatore gcc, si lamentano se un file contiene la firma UTF-8. Molti programmi Windows utilizzano invece la firma. E il tentativo di rilevare la codifica di un flusso di byte non sempre funziona.

Come verificare se il tuo progetto ha codifica UTF-8 o meno

UTF-8 non è ancora universale e gli ingegneri del software e gli scienziati dei dati spesso devono affrontare problemi di codifica dei flussi di testo. A volte si suppone che UTF-8 venga utilizzato nel progetto, tuttavia viene utilizzato un altro ecndoing. Esistono diversi strumenti per rilevare la codifica del file:

- Alcuni strumenti CMD, come lo strumento da riga di comando Linux ' file ' o `powershell` ;
- Pacchetto Python "chardet"
- Notepad ++ come forse lo strumento più popolare per il controllo manuale.

Examples

Come convertire un array di byte di dati UTF-8 in una stringa Unicode in Python

```
def make_unicode(data):
    if type(data) != unicode:
        data =  data.decode('utf-8')
        return data
    else:
        return data
```

Come cambiare la codifica predefinita del server in UTF-8

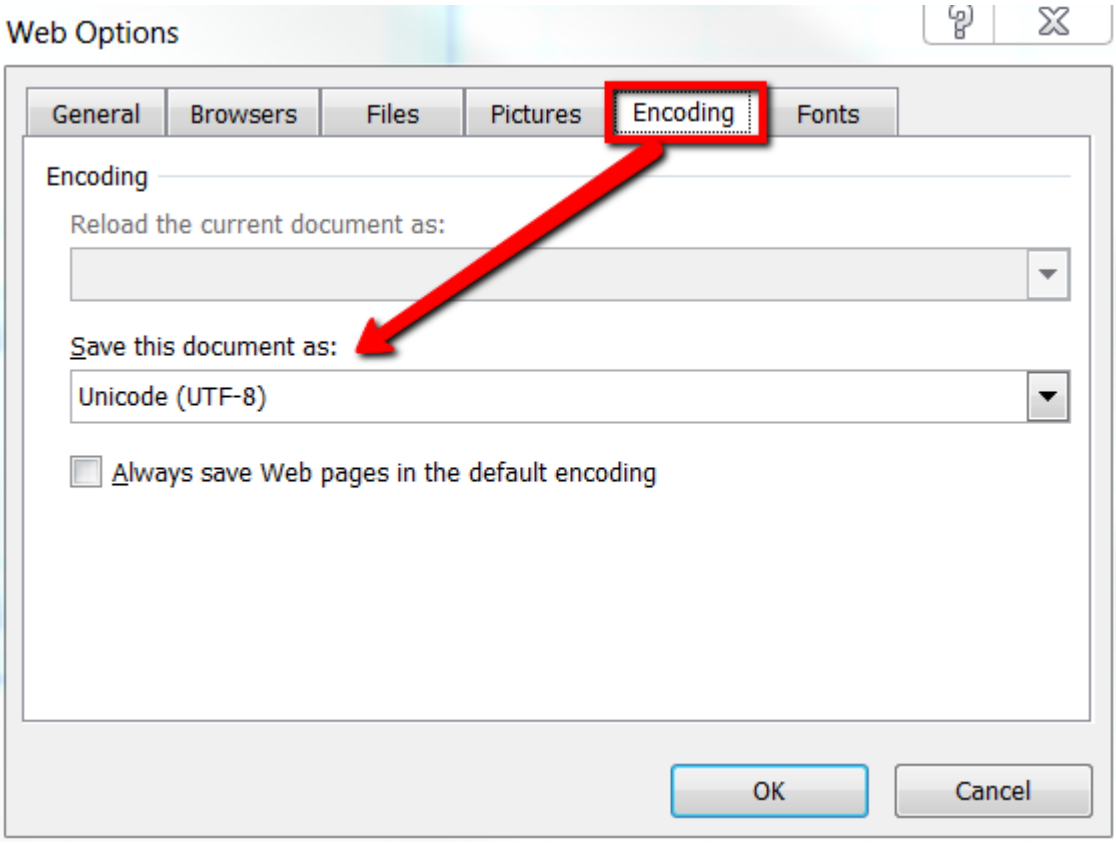
A volte gli utenti di altre regioni che parlano l'inglese hanno problemi con la codifica mentre per esempio programmano un progetto php. Può essere che il server abbia un'altra codifica e quindi UTF-8 e se qualcuno vuole creare un progetto php in UTF-8 su questo server, il suo testo potrebbe essere visualizzato in modo errato.

Esempio: può essere che sul tuo server la codifica predefinita sia Windows-1251 - quindi dovresti eliminare `AddDefaultCharset windows-1251` dal **file del** server **.htaccess** e scrivere `AddDefaultCharset utf-8` .

Per verificare, quale codifica ha il tuo server, non impostare il `<META charset>` e attivare il `"automatic encoding detection"` nel tuo browser.

Salva un file Excel in UTF-8

Excel -> Salva come -> Salva come -> "Valore separato da virgola (* .csv)" E Strumenti (da sinistra a pulsante Salva) -> Opzioni Web -> Codifica -> Salva questo documento come -> Unicode (UTF-8)



Leggi UTF-8 come un modo di codifica di Unicode online:
https://riptutorial.com/it/unicode/topic/6035/utf-8-come-un-modo-di-codifica-di-unicode

Titoli di coda

S. No	Capitoli	Contributors
1	Iniziare con Unicode	Community, DPenner1
2	I caratteri possono essere costituiti da più punti di codice	roeland
3	Il testo inglese non è solo ASCII	roeland
4	UTF-8 come un modo di codifica di Unicode	R. Martinho Fernandes, vlad.rad