EBook Gratis

# APRENDIZAJE unicode

Free unaffiliated eBook created from
**Stack Overflow contributors.**

#unicode

# Tabla de contenido


Acerca de ..... 1

Capítulo 1: Empezando con Unicode ..... 2

Observaciones ..... 2

Versiones ..... 2

Examples ..... 3

Instalación o configuración ..... 3

Capítulo 2: El texto en inglés no es solo ASCII ..... 4

Observaciones ..... 4

Examples ..... 4

Diacríticos ..... 4

Emoji ..... 4

Puntuación ..... 4

Simbolos especiales ..... 5

Capítulo 3: Los personajes pueden constar de múltiples puntos de código ..... 6

Observaciones ..... 6

Examples ..... 6

Diacríticos ..... 6

formas combinadas ..... 6

Texto de zalgo ..... 6

Emoji y banderas ..... 7

Capítulo 4: UTF-8 como una forma de codificación de Unicode ..... 8

Observaciones ..... 8

Examples ..... 9

Cómo convertir una matriz de bytes de datos UTF-8 en una cadena Unicode en Python ..... 9

Cómo cambiar la codificación predeterminada del servidor a UTF-8 ..... 9

Guardar un archivo de Excel en UTF-8 ..... 9

Creditos ..... 11

# Acerca de

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: unicode

It is an unofficial and free unicode ebook created for educational purposes. All the content is extracted from Stack Overflow Documentation, which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official unicode.



Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

# Capítulo 1: Empezando con Unicode

## Observaciones

El estándar de Unicode es un conjunto de caracteres internacional estandarizado. Intenta asignar caracteres y símbolos de cada sistema de escritura a un número único. Con cada nueva versión principal, se agregan caracteres adicionales al Estándar para lograr este objetivo. Al proporcionar un conjunto de caracteres unificado para todos los sistemas de escritura, la información de texto se puede intercambiar en un formato Unicode independiente de cualquier plataforma dada.

El estándar de Unicode también contiene datos de propiedad sobre los caracteres y define algoritmos sobre cómo manipular correctamente los caracteres. Por ejemplo, estos algoritmos proporcionan el método correcto para buscar y mostrar texto Unicode.

## Versiones

| Versión | Fecha de lanzamiento |
|---|---|
| 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 |

| Versión | Fecha de lanzamiento |
| --- | --- |
| 8.0.0 | 2015-06-17 |
| 9.0.0 | 2016-06-21 |

# Examples

## Instalación o configuración

Instrucciones detalladas sobre cómo configurar o instalar Unicode.

Lea Empezando con Unicode en línea: https://riptutorial.com/es/unicode/topic/3188/empezando-con-unicode

# Capítulo 2: El texto en inglés no es solo ASCII

## Observaciones

Una suposición que aparece regularmente es que cuando se trata solo de texto en inglés, es poco probable que encuentre caracteres fuera del conjunto de caracteres ASCII. Para evitar problemas con el manejo correcto de Unicode, las personas tienen la tentación de hacer cosas como eliminar caracteres que no son ASCII o eliminar cualquier acento en las letras.

Estos ejemplos muestran que esta suposición es incorrecta, e incluso para el texto en inglés, debe tener cuidado de manejar los caracteres Unicode correctamente.

## Examples

### Diacríticos

El texto en inglés tiene los diacríticos ocasionales.

- Palabras de préstamo, como née, café, plato principal.
- Nombres, como Noël y Chloë
- Nombres de lugares, como Montreal y Quebec

### Emoji

Los emoji son muy populares en las redes sociales en estos días.

- : `U+2603` - MUÑECO DE NIEVE
- : `U+01F600` - GRINNING FACE
- : `U+01F42A` - CAMEL DROMEDARIO

Tenga en cuenta que la mayoría de los emoji están fuera del plano multilingüe básico. Una gran cantidad de nuevas adiciones consisten en más de un punto de código:

- : Una bandera se define como un par de "letras indicadoras de símbolos regionales"
- : Este es un emoji más un modificador de tono de piel:  +
- o : Windows 10 le permite especificar si un emoji es de color o blanco / negro agregando un selector de variación ( `U+FE0E` o `U+FE0F` )

### Puntuación

Casi todo el texto escrito tiene signos de puntuación que están fuera del conjunto de caracteres ASCII:

- guiones: el guión en -, y el guión em -

- Las comillas: "comillas" en lugar de "comillas"
- Los puntos suspensivos ...

## Simbolos especiales

Hay algunos símbolos comunes en uso:

- Copyright © y marcas registradas ® ™
- fracciones como ¼
- superíndices. Por ejemplo, una taquigrafía para metros cuadrados es m².

Lea El texto en inglés no es solo ASCII en línea: https://riptutorial.com/es/unicode/topic/5198/el-texto-en-ingles-no-es-solo-ascii

# Capítulo 3: Los personajes pueden constar de múltiples puntos de código

## Observaciones

Un punto de código Unicode, lo que los programadores a menudo piensan en un personaje, a menudo corresponde a lo que el usuario piensa que es un personaje. A veces, sin embargo, un "carácter" se compone de múltiples puntos de código, como muestran los ejemplos anteriores.

Esto significa que las operaciones como cortar una cadena o obtener un carácter en un índice determinado pueden no funcionar como se espera. Por ejemplo, el $4^o$ carácter de la cadena `"Café "` es `'e'` (sin el acento). Del mismo modo, cortar la cuerda a la longitud 4 eliminará el acento.

El término técnico para tal grupo de puntos de código es un *grupo de grafemas* . Ver UAX # 29: Segmentación de texto Unicode

## Examples

### Diacríticos

Una letra con un diacrítico se puede representar con la letra y una letra modificadora de combinación. Normalmente piensas en `é` como un personaje, pero en realidad son 2 puntos de código:

- `U+0065` - LETRA PEQUEÑA LATINA E
- `U+0301` - COMBINANDO ACENTO AGUDO

De manera similar, `ç` = `c` + `¸` , y `å` = `a` + `°`

## formas combinadas

Para complicar las cosas, a menudo hay un punto de código para la forma compuesta también:

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

Aunque estas cuerdas tienen el mismo aspecto, no son iguales y ni siquiera tienen la misma longitud (5 y 4 respectivamente).

### Texto de zalgo

Hay una cosa llamada Texto de Zalgo que empuja esto al extremo. Aquí está el primer grupo de grafemas del ejemplo. Consta de 15 puntos de código: la letra latina `H` y 14 marcas combinadas.

Aunque esto no aparece en el texto normal, muestra que un "carácter" realmente puede consistir en un número arbitrario de puntos de código

## Emoji y banderas

Una gran cantidad de emoji consisten en más de un punto de código.

- : Una bandera se define como un par de "letras indicadoras de símbolos regionales" ( + )
- : Algunos emoji pueden ir seguidos de un modificador de tono de piel:  + 
- o : Windows 10 le permite especificar si un emoji es de color o blanco / negro agregando un selector de variación ( `U+FE0E` o `U+FE0F` )
- : una familia. Codificado uniendo el emoji para niño, niña, mujer y hombre (, , , ) junto con uniones de ancho cero ( `U+200D` ). En las plataformas que lo soportan, esto se representa como un emoji de una familia con dos hijos.

Lea Los personajes pueden constar de múltiples puntos de código en línea: https://riptutorial.com/es/unicode/topic/6485/los-personajes-pueden-constar-de-multiples-puntos-de-codigo

# Capítulo 4: UTF-8 como una forma de codificación de Unicode

## Observaciones

**¿Qué es UTF-8** ?

UTF-8 es una codificación, que es de longitud variable y utiliza unidades de código de 8 bits, por eso UTF- **8** . En Internet, UTF-8 es una codificación dominante (antes de 2008 ASCII, que también puede manejar cualquier punto de código Unicode).

**¿Es UTF-8 lo mismo que Unicode?**

"Unicode" no es una codificación, es un conjunto de caracteres codificados, es decir, un conjunto de caracteres y una asignación entre los caracteres y los puntos de código entero que los representan. Pero una gran cantidad de documentación lo utiliza para referirse a las *codificaciones* . En Windows, por ejemplo, el término Unicode se usa para referirse a UTF-16.

UTF-8 es solo una de las formas de codificar Unicode y, como codificación, convierte las secuencias de bytes en secuencias de caracteres y viceversa. UTF-16 y -32 son otros formatos de transformación Unicode.

**Lista de materiales de UTF-8**

Los tres pueden tener una marca de orden de bytes específica, que al ser un número mágico señala varias cosas importantes para un programa (por ejemplo, Notepad ++); por ejemplo, el hecho de que el flujo de texto importado es Unicode; También ayuda a detectar el arte de Unicode utilizado para este flujo. Sin embargo, el consorcio Unicode recomienda almacenar UTF-8 sin ninguna firma. Algún software, por ejemplo, el compilador gcc se queja si un archivo contiene la firma UTF-8. Una gran cantidad de programas de Windows por otro lado usan la firma. Y tratar de detectar la codificación de un flujo de bytes no siempre funciona.

**Cómo verificar si su proyecto tiene codificación UTF-8 o no**

UTF-8 aún no es universal, y los ingenieros de software y los científicos de datos a menudo enfrentan problemas de codificación de flujos de texto. A veces se supone que se usa UTF-8 en el proyecto, sin embargo, se está utilizando otro proceso de creación. Existen varias herramientas para detectar la codificación del archivo:

- Algunas herramientas de CMD, como la herramienta de línea de comandos de Linux ' archivo ' o `powershell`
- Paquete Python "chardet"
- Notepad ++ como la herramienta más popular para la comprobación manual.

# Examples

## Cómo convertir una matriz de bytes de datos UTF-8 en una cadena Unicode en Python

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

## Cómo cambiar la codificación predeterminada del servidor a UTF-8

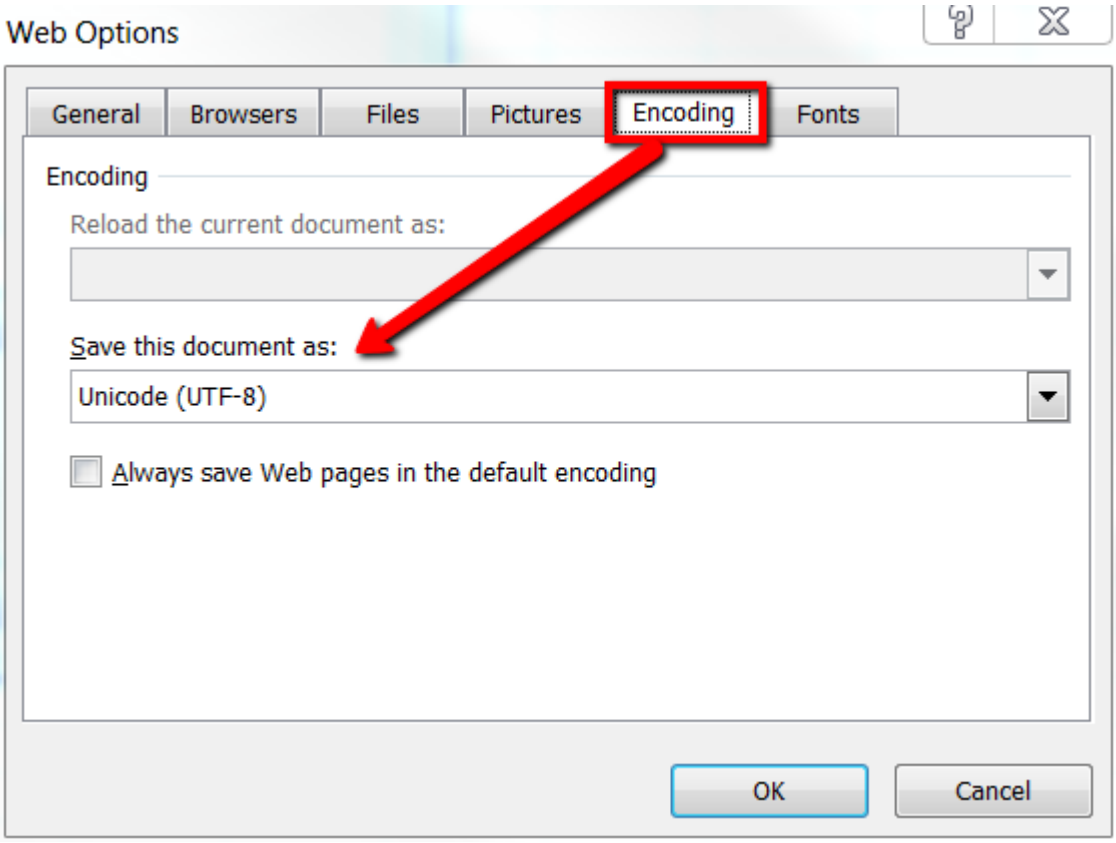
A veces, los usuarios de otras regiones que no son de habla inglesa tienen problemas con la codificación mientras que, por ejemplo, programan un proyecto php. Puede ser que el servidor tenga otra codificación, luego UTF-8, y si alguien quiere crear un proyecto php en UTF-8 en este servidor, su texto puede aparecer incorrecto.

Ejemplo: puede ser que en su servidor la codificación predeterminada sea Windows-1251; luego, debe eliminar `AddDefaultCharset windows-1251` del archivo del servidor **.htaccess** y escribir `AddDefaultCharset utf-8` .

Para verificar qué codificación tiene su servidor, no configure la etiqueta `<META charset>` y active la `"automatic encoding detection"` en su navegador.

## Guardar un archivo de Excel en UTF-8

Excel -> Guardar como -> Guardar como tipo -> "Valor separado por comas (* .csv)" Y Herramientas (a la izquierda para guardar) -> Opciones web -> Codificar -> Guardar este documento como -> Unicode (UTF-8) )




Lea UTF-8 como una forma de codificación de Unicode en línea:
https://riptutorial.com/es/unicode/topic/6035/utf-8-como-una-forma-de-codificacion-de-unicode

# Creditos

| S. No | Capítulos | Contributors |
|---|---|---|
| 1 | Empezando con Unicode | Community, DPenner1 |
| 2 | El texto en inglés no es solo ASCII | roeland |
| 3 | Los personajes pueden constar de múltiples puntos de código | roeland |
| 4 | UTF-8 como una forma de codificación de Unicode | R. Martinho Fernandes, vlad.rad |