FREE eBook

LEARNING unicode

Free unaffiliated eBook created from
Stack Overflow contributors.

#unicode

Table of Contents

About

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

Chapter 1: Getting started with unicode

Remarks

The Unicode Standard is an international standardized character set. It attempts to assign characters and symbols from every writing system a unique number. With every major new version, additional characters are added to the Standard to achieve this goal. In providing a unified character set for all writing systems, text information can be exchanged in a Unicode format independent of any given platform.

The Unicode Standard also contains property data on the characters, and defines algorithms on how to properly manipulate characters. For example, these algorithms provide the correct method to search and display Unicode text.

Versions

Version	Release Date
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

Version	Release Date
8.0.0	2015-06-17
9.0.0	2016-06-21

Examples

Installation or Setup

Detailed instructions on getting unicode set up or installed.

Read Getting started with unicode online: https://riptutorial.com/unicode/topic/3188/getting-started-with-unicode

Chapter 2: Characters can consist of multiple code points

Remarks

An Unicode code point, what programmers often think of one character, often corresponds to what the user thinks is one character. Sometimes however a “character” is made up of multiple code points, as the examples above show.

This means that operations like slicing a string, or getting a character at a given index may not work as expected. For instance the 4th character of the string `"Café "` is `'e'` (without the accent). Similarly, clipping the string to length 4 will remove the accent.

The technical term for such a group of code points is a *grapheme cluster*. See UAX #29: Unicode Text Segmentation

Examples

Diacritics

A letter with a diacritic may be represented with the letter, and a combining modifier letter. You normally think of `é` as one character, but it's really 2 code points:

- `U+0065` — LATIN SMALL LETTER E
- `U+0301` — COMBINING ACUTE ACCENT

Similarly `ç` = `c` + `¸`, and `å` = `a` + `°`

combined forms

To complicate matters, there is often a code point for the composed form as well:

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

Although these strings look the same, they are not equal, and they don't even have the same length (5 and 4 respectively).

Zalgo Text

There is this thing called Zalgo Text which pushes this to the extreme. Here is the first grapheme cluster of the example. It consists of 15 code points: the Latin letter `H` and 14 combining marks.

Although this doesn't show up in normal text, it shows that a “character” really can consist of an arbitrary number of code points

Emoji and flags

A lot of emoji consist of more than one code point.

- : A flag is defined as a pair of "regional symbol indicator letters" (+)
- : Some emoji may be followed by a skin tone modifier: +
- or : Windows 10 allows you to specify if an emoji is colored or black/white by appending a variation selector (`U+FE0E` or `U+FE0F`)
- : a family. Encoded by joining the emoji for boy, girl, woman and man (, , ,) together with zero-width joiners (`U+200D`). On platforms which support it, this is rendered as an emoji of a family with two kids.

Read Characters can consist of multiple code points online:
https://riptutorial.com/unicode/topic/6485/characters-can-consist-of-multiple-code-points

Chapter 3: English text is not ASCII only

Remarks

An assumption which pops up regularly is that when dealing with English text only, it's unlikely to encounter characters outside the ASCII character set. To avoid problems with handling Unicode correctly, people are tempted to do things like stripping non-ASCII characters, or removing any accents on letters.

These examples show this assumption is wrong, and even for English text you should take care to handle Unicode characters correctly.

Examples

Diacritics

English text has the occasional diacritics.

- Loan words, like née, café, entrée
- Names, like Noël and Chloë
- Place names, like Montréal and Québec

Emoji

Emoji are quite popular with social media these days.

- : `U+2603` — SNOWMAN
- : `U+01F600` — GRINNING FACE
- : `U+01F42A` — DROMEDARY CAMEL

Note that most emoji are outside the Basic Multilingual Plane. A lot of newer additions consist of more than one code point:

- : A flag is defined as a pair of "regional symbol indicator letters"
- : This is an emoji plus a skin tone modifier: +
- or : Windows 10 allows you to specify if an emoji is colored or black/white by appending a variation selector (`U+FE0E` or `U+FE0F`)

Punctuation

Almost all written text has punctuation marks which are outside the ASCII character set:

- dashes: the en dash –, and the em dash —
- Quotation marks: “quotes” rather than "quotes"
- The ellipsis…

Special symbols

There are a few common symbols in use:

- copyright sign ©, and trademark signs ® ™
- fractions like ¼
- superscripts. For instance, a shorthand for square meters is m².

Read English text is not ASCII only online: https://riptutorial.com/unicode/topic/5198/english-text-is-not-ascii-only

Chapter 4: UTF-8 as an encoding way of Unicode

Remarks

What is UTF-8?

UTF-8 is an encoding, which is variable-length and uses 8-bit code units - that's why UTF-**8**. In the internet UTF-8 is dominant encoding (before 2008 ASCII was, ehich also can handle any Unicode code point.).

Is UTF-8 the same as Unicode?

"Unicode" isn't an encoding - it is a coded character set - i.e. a set of characters and a mapping between the characters and integer code points representing them. But a lot of documentation uses it to refer to *encodings*. On Windows, for example, the term Unicode is used to refer to UTF-16.

UTF-8 is only one of the ways to encode Unicode and as an encoding it converts the sequences of bytes to sequences of characters and vice versa. UTF-16 and -32 are other Unicode transformation formats.

BOM of UTF-8

All three mayhave a specific Byte Order Marks, which being a magic number signals several important things to a program (for example, Notepad++) - for example, the fact, that the imported text stream is Unicode; also it helps to detect the art of Unicode used for this stream. However the Unicode consortium recommends storing UTF-8 without any signature. Some software, for example gcc compiler complains if a file contains the UTF-8 signature. A lot of Windows programs on the other hand use the signature. And trying to detect the encoding of a stream of bytes don't always work.

How to check if your project has UTF-8 encoding or not

UTF-8 is yet not universal, and software engineers and data scientists often face problem of encoding of text streams. Sometimes UTF-8 is supposed to be used in the project, however another ecndoing is being used. There are several tools to detect the encoding of the file:

- Some CMD tools, like Linux command-line tool 'file' or `powershell`;
- Python package "chardet"
- Notepad++ as maybe the most popular tool for manual check.

Examples

How to convert a byte array of UTF-8 data to a Unicode string in Python

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

How to change the default encoding of the server to UTF-8

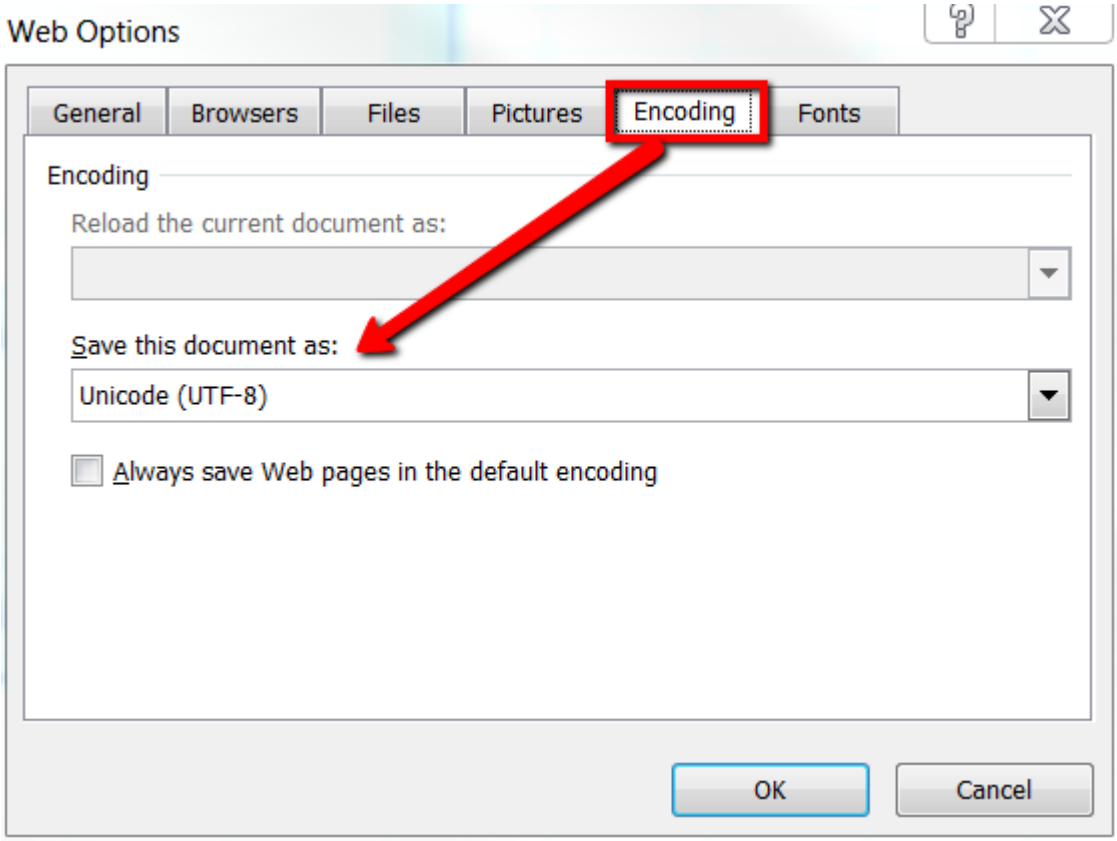
Sometimes users from other regions than English-speaking have problems with encoding while for example programming a php project. It can be, that the server has another encoding then UTF-8, and if someone want to create a php project in UTF-8 on this server, his text might be shown incorrect.

Example: it can be that on your server default encoding is Windows-1251 - then you should delete the `AddDefaultCharset windows-1251` from the **.htaccess** server file and write `AddDefaultCharset utf-8` .

To check, which encoding does your server have, don't set the `<META charset>` tag and activate `"automatic encoding detection"` in your browser.

Save an Excel file in UTF-8

Excel -> Save as -> Save as type -> "Comma separated value (*.csv)" AND Tools (left to Save button) -> Web options -> Encoding -> Save this document as -> Unicode (UTF-8)



Read UTF-8 as an encoding way of Unicode online: https://riptutorial.com/unicode/topic/6035/utf-8-as-an-encoding-way-of-unicode

Credits

S. No	Chapters	Contributors
1	Getting started with unicode	Community, DPenner1
2	Characters can consist of multiple code points	roeland
3	English text is not ASCII only	roeland
4	UTF-8 as an encoding way of Unicode	R. Martinho Fernandes, vlad.rad