@omiita
投稿日 2019年11月25日　更新日 2019年11月25日

【図解】【3分解説】UnicodeとUTF-8の違い！【今さら聞けない】

プログラミング, UTF-8, 初心者, 文字コード, unicode

【図解】【3分解説】UnicodeとUTF-8の違い！【今さら聞けない】

はじめに

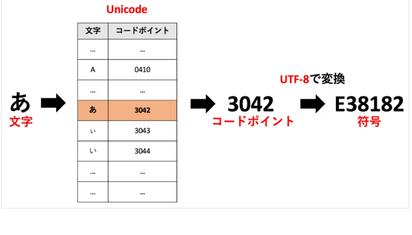
UnicodeとUTF-8の違いについて理解が曖昧だったので、調べました。コード書いてたまに巡り会いますよね。
調べたら説明が長いサイトが多く、予想以上に時間がかかってしまったのでここでは直感的に説明。

読んで分かりやすかったり少しでも何か学べたと思えたら **いいね** や **コメント** をもらえるとこれからの励みになります！よろしくお願いします！

~流れ~
1. さっそくざっくり説明
2. まとめ
3. もう少し時間がある方へ
4. 雑談
5. 参考

さっそくざっくり説明

あくまで私がしっくりくる説明です。UnicodeとUTF-8の私のイメージを図で表すと下図。



【説明】
パソコンが文字を理解するまでには **文字→[Unicodeを参照]→コードポイント→[UTF-8による変換]→符号→[パソコン理解]** という流れ。

- **Unicode**: 世界中のあらゆる文字たちにとりあえず16進数の数字(**コードポイント**)を割り振った表。つまり、**「文字 → コードポイント」に変換する表**のようなもの。

> Ex.)「あ」のコードポイントは「3042」

- **UTF-8**: Unicodeで割り振ったコードポイントをパソコンがわかるように別の16進数の数字(**符号**)に変換する方法の1つ。変換方法の違いでUTF-16とかUTF-32とかの仲間がいます。つまり、**「コードポイント → 符号」に変換する方法** の1つ。

> Ex.) コードポイント「3042」をUTF-8で符号に変換すると「E38182」(16進数)

これでざっくりと

- **Unicodeが文字からコードポイントに変換するための表**のようなもの。
- **UTF-8がコードポイントから符号に変換する方式**

ということがわかって頂けたかと思います。以上です。
(Unicodeのコードポイントをそのままパソコンに理解させるよりもUTF-8などによる符号を使う方がバイト削減できるため、UTF-8で符号化している。詳細は後述の雑談)

まとめ

言葉	説明
コードポイント	Unicodeがあらゆる文字につけた数字
符号	パソコンが理解してくれる数字
Unicode	**文字→コードポイントしてくれる表**みたいなもの
UTF-8	**コードポイント→符号してくれる方式**の1つ

UTF-16のことをUnicodeと記しているソフトウエア(Windowsのメモ帳など)もありますのでUnicodeとあったらそれはUTF-16を使って変換したものなのだな、というふうに理解してください。 そうなってしまっている理由はこちらで解説されていました。

これでUnicodeとUTF-8の違いはバッチリですね！おわり。

読んで分かりやすかったり少しでも何か学べたと思えたら **いいね** や **コメント** をもらえるとこれからの励みになります！

もう少し時間がある方へ

手計算で文字をUTF-8での符号まで計算してみましょう。
理解が一気に深まります。手順は以下。

1. **文字のコードポイントをUnicodeから見つけてくる。**
2. **コードポイントをUTF-8の方式で変換してみる。**

Omiitaの「お」をUTF-8による符号まで変換してみます。

1. **文字「お」のコードポイントをUnicodeから見つけてくる。**

> 「お」のコードポイント:304A

2. **コードポイント「304A」をUTF-8の方式で変換してみる。**
2.1 まず「304A」を2進数に変えます。
2.2 (コードポイントが0800 ... FFFFの間にある今回の場合)、先頭から4, 6, 6ビットに分けます。
2.3 分けたものにそれぞれ先頭からE0,80,80を足せば、終了です。

> [2.1] 304A => 0011000001001010
> [2.2] 0011000001001010 => 0011 / 000001 / 001010 => 03 / 01 / 0A (16進数)
> [2.3] 03010A + E08080 = **E3818A**

これで「お」はUTF-8で「E3818A」であると計算できました。
あとはこちらのサイトでCtrl+Fで「E3818A」を検索すると「お」であることがわかります。
(10進数で示すなら「お」=> 「E3 81 8A」 => 「227 129 138」となります。)

雑談

- 厳密にはUnicodeは「符号化文字集合体」、UTF-8などは「文字符号化方式」と日本語で言われています。
- UTF-8ではアルファベットは1バイトでひらがなは3バイトです。
- **コードポイントをそのまま符号として使う方式はUTF-32** と呼ばれています。
- UTF-32のようにコードポイントをそのまま使おうとすると世の中の文字は大量にあるので、あらかじめコードポイントを4バイト長にしなくてはならないので**符号が全て4バイトになってしまい非効率です。**
- 符号化文字集合体(Unicode)を文字符号化方式(UTF-8など)で符号化する意義についてはこちらの質問でまさに議論されています。

読んで分かりやすかったり少しでも何か学べたと思えたら **いいね** や **コメント** をもらえるとこれからの励みになります！よろしくお願いします！

参考

- Unicode一覧 3000-3FFF
  ひらがなが入っているUnicodeの箇所(Wikipedia)
- 新人さんに知ってほしい「文字コードのお話」
  文字コードをしっかりと知りたい方へのQiita記事
- UTF-8
  UTF-8の説明(Wikipedia)