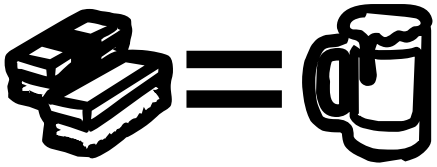


とみたまさひろ

MyNA会

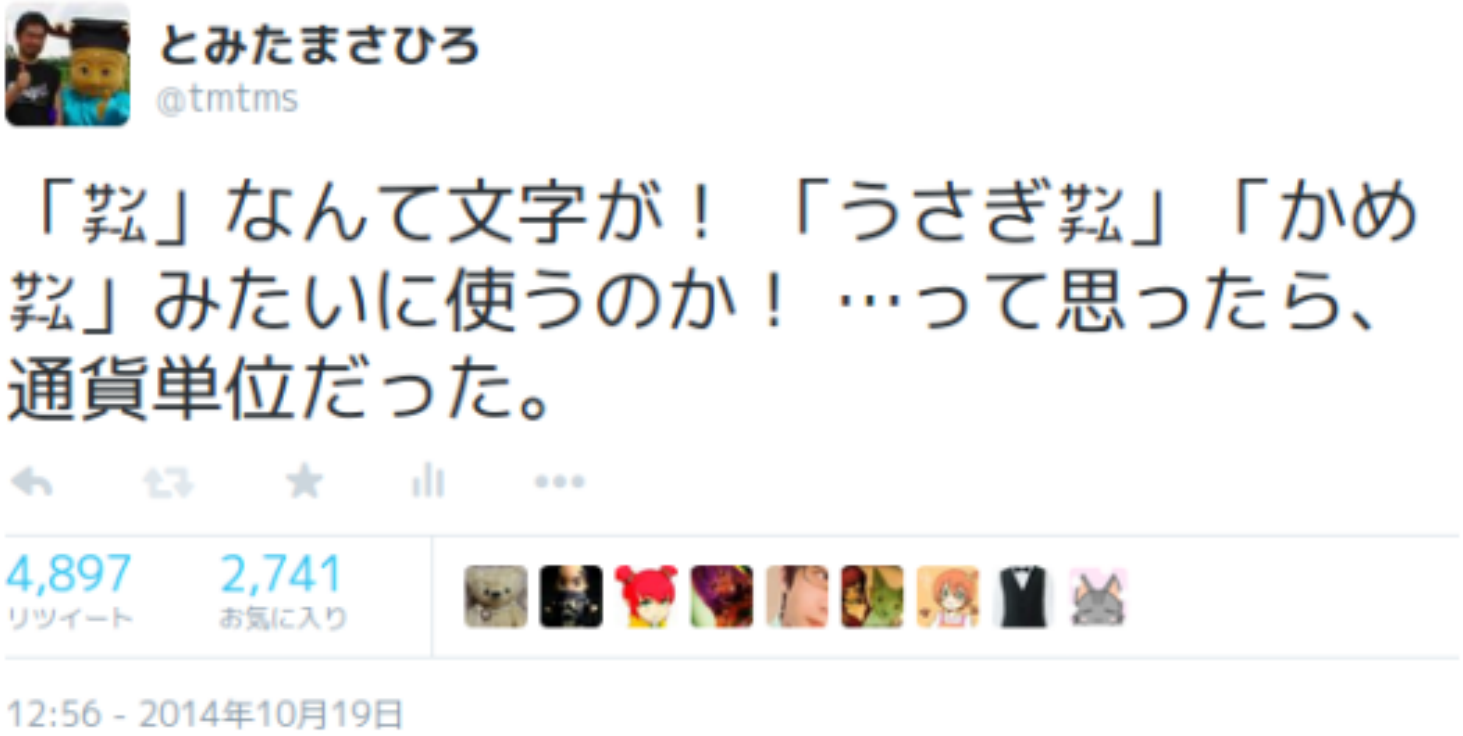
2015/04/22

自己紹介

- とみた まさひろ
  - http://**tmtms**.hatenablog.com
  - http://twitter.com/**tmtms**
  - https://github.com/**tmtm**
- MySQL 3.21 に日本語charsetを追加
- MySQLのRubyバインディング作成

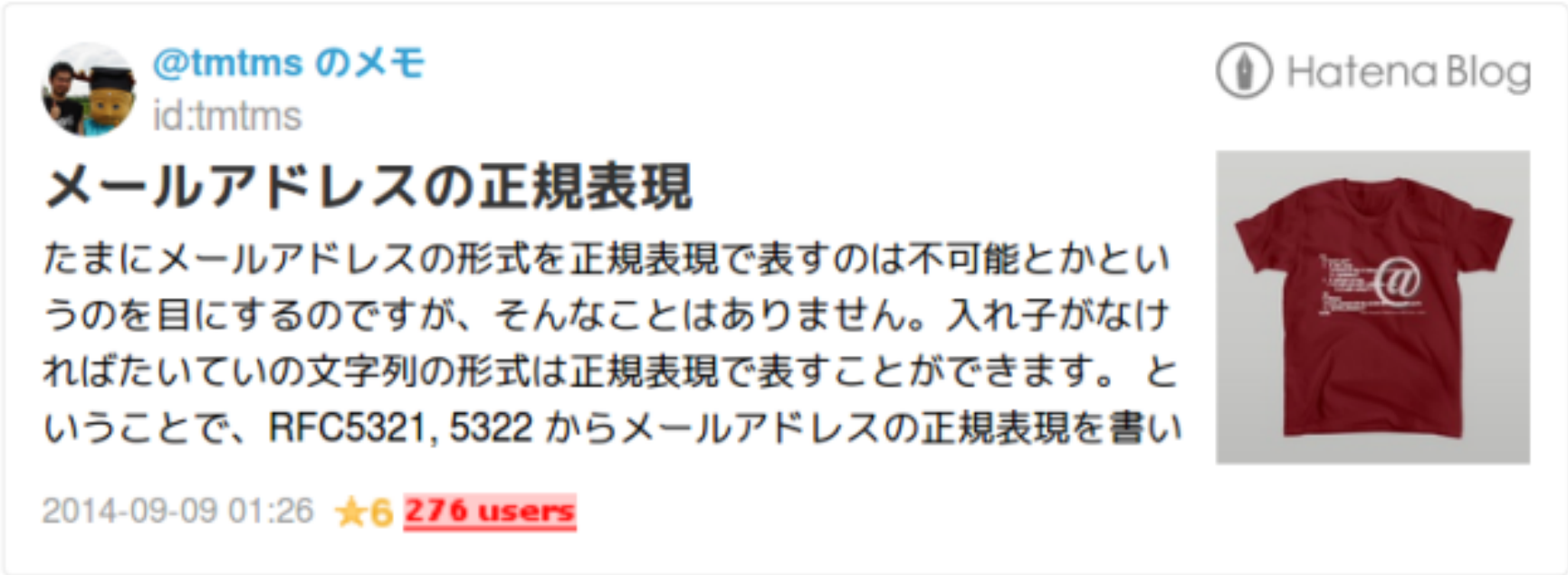
自己紹介

- もっともRTされたツイート



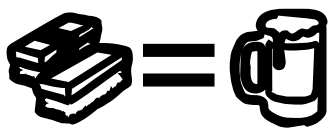
自己紹介

- もっともブクマされたブログ



自己紹介

- 長野県北部在住
- 日本MySQLユーザ会代表
- 名ばかり代表
- 「たまには何かしゃべれや (#°Д°)ｺﾞﾙｧ!!」
  と言われたのでしゃべります

 = 問題

MySQL的には📚と🍺は同じ

とみたまさひろ
@tmtms

あれ？ MySQL の utf8mb4 charset って、4バイト文字同士を比較すると同じ文字扱いされる？
SELECT '🍣'='🍺' → 1
MySQL的には寿司とビールは同じ扱い。

243 リツイート　159 お気に入り

16:10 - 2014年12月22日

ちなみに  と  も(ry

PostgreSQLなら問題ないらしい

http://soudai1025.blogspot.jp/2015/03/postgresqlunicode-6.html

何故？

kamipo++

- utf8_unicode_ci に対する日本の開発者の見解
  http://blog.kamipo.net/entry/2015/03/08/145045
- MySQL と Unicode Collation Algorithm (UCA)
  http://blog.kamipo.net/entry/2015/03/17/103457
- MySQL と寿司ビール問題
  http://blog.kamipo.net/entry/2015/03/23/093052

MySQLの文字は Charset と Collation がある

Charset

いわゆる文字コード

文字のバイト表現

Charset: utf8mb4

- 「A」= 41
- 「あ」= E3 81 82
- 「  」= F0 9F 8D A3
- 「  」= F0 9F 8D BA

Collation

文字の照合規則・照合順序

Collation 一覧

```
mysql> show collation;
+---------------------------+----------+-----+---------+----------+---------+
| Collation                 | Charset  | Id  | Default | Compiled | Sortlen |
+---------------------------+----------+-----+---------+----------+---------+
| big5_chinese_ci           | big5     |   1 | Yes     | Yes      |       1 |
| big5_bin                  | big5     |  84 |         | Yes      |       1 |
| dec8_swedish_ci           | dec8     |   3 | Yes     | Yes      |       1 |
| dec8_bin                  | dec8     |  69 |         | Yes      |       1 |
| cp850_general_ci          | cp850    |   4 | Yes     | Yes      |       1 |
| cp850_bin                 | cp850    |  80 |         | Yes      |       1 |
| hp8_english_ci            | hp8      |   6 | Yes     | Yes      |       1 |
| hp8_bin                   | hp8      |  72 |         | Yes      |       1 |
| koi8r_general_ci          | koi8r    |   7 | Yes     | Yes      |       1 |
| koi8r_bin                 | koi8r    |  74 |         | Yes      |       1 |
| latin1_german1_ci         | latin1   |   5 |         | Yes      |       1 |
| latin1_swedish_ci         | latin1   |   8 | Yes     | Yes      |       1 |
| latin1_danish_ci          | latin1   |  15 |         | Yes      |       1 |
| latin1_german2_ci         | latin1   |  31 |         | Yes      |       2 |
| latin1_bin                | latin1   |  47 |         | Yes      |       1 |
| latin1_general_ci         | latin1   |  48 |         | Yes      |       1 |
| latin1_general_cs         | latin1   |  49 |         | Yes      |       1 |
```

Charset 毎に Collation がある

utf8mb4 の Collation

全部で16個

```
mysql> show collation like 'utf8mb4%';
+-------------------------+---------+-----+---------+----------+---------+
| Collation               | Charset | Id  | Default | Compiled | Sortlen |
+-------------------------+---------+-----+---------+----------+---------+
| utf8mb4_general_ci      | utf8mb4 |  45 | Yes     | Yes      |       1 |
| utf8mb4_bin             | utf8mb4 |  46 |         | Yes      |       1 |
| utf8mb4_unicode_ci      | utf8mb4 | 224 |         | Yes      |       8 |
| utf8mb4_icelandic_ci    | utf8mb4 | 225 |         | Yes      |       8 |
| utf8mb4_latvian_ci      | utf8mb4 | 226 |         | Yes      |       8 |
| utf8mb4_romanian_ci     | utf8mb4 | 227 |         | Yes      |       8 |
| utf8mb4_slovenian_ci    | utf8mb4 | 228 |         | Yes      |       8 |
| utf8mb4_polish_ci       | utf8mb4 | 229 |         | Yes      |       8 |
| utf8mb4_estonian_ci     | utf8mb4 | 230 |         | Yes      |       8 |
| utf8mb4_spanish_ci      | utf8mb4 | 231 |         | Yes      |       8 |
| utf8mb4_swedish_ci      | utf8mb4 | 232 |         | Yes      |       8 |
```

utf8mb4 の Collation

utf8mb4_turkish_ci	utf8mb4	233		Yes	8
utf8mb4_czech_ci	utf8mb4	234		Yes	8
utf8mb4_danish_ci	utf8mb4	235		Yes	8
utf8mb4_lithuanian_ci	utf8mb4	236		Yes	8
utf8mb4_slovak_ci	utf8mb4	237		Yes	8
utf8mb4_spanish2_ci	utf8mb4	238		Yes	8
utf8mb4_roman_ci	utf8mb4	239		Yes	8
utf8mb4_persian_ci	utf8mb4	240		Yes	8
utf8mb4_esperanto_ci	utf8mb4	241		Yes	8
utf8mb4_hungarian_ci	utf8mb4	242		Yes	8
utf8mb4_sinhala_ci	utf8mb4	243		Yes	8
utf8mb4_german2_ci	utf8mb4	244		Yes	8
utf8mb4_croatian_ci	utf8mb4	245		Yes	8
utf8mb4_unicode_520_ci	utf8mb4	246		Yes	8
utf8mb4_vietnamese_ci	utf8mb4	247		Yes	8

utf8mb4 の Collation

- utf8mb4_general_ci
- utf8mb4_bin
- utf8mb4_unicode_ci
- utf8mb4_unicode_520_ci
- utf8mb4_言語_ci
  (utf8m4_japanese_ci は無い)

utf8mb4_general_ci

- utf8mb4 charset のデフォルト collation
- ASCII大文字小文字を区別しない(A=a)
- 絵文字を**区別しない**( = )

utf8mb4_bin

- varchar(99) binary
- 全文字を**区別する**(A≠a, 📚≠🍺)
- PostgreSQLと同じならこれでいい

utf8mb4_unicode_ci

- Unicode Collation Algorithm 4.0.0

  http://www.unicode.org/reports/tr10/
  http://dev.mysql.com/doc/refman/5.6/en/charset-unicode-sets.html

- ASCII大文字小文字を区別しない(A=a)
- 絵文字を**区別しない**(📚=🍺)
- ひらがな、カタカナ、濁点有無、全角、半角を区別しない((は=ば=ぱ=ハ=バ=パ=ﾊ)

utf8mb4_unicode_520_ci

- Unicode Collation Algorithm 5.2.0
- ASCII大文字小文字を区別しない(A=a)
- 絵文字を**区別する**(📚≠🍺)
- ひらがな、カタカナ、濁点有無、全角、半角を区別しない((は=ば=ぱ=ハ=バ=パ=ﾊ)

ハハ=パパ=ババ 問題

誰得

utf8mb4_*_ci

Collation	A : a	 :	は : ぱ
general	=	=	≠
bin	≠	≠	≠
unicode	=	=	=
unicode_520	=	≠	=

ぼくらが本当に欲しかったもの

Collation	A : a		は : ぱ
general	=	=	≠
bin	≠	≠	≠
unicode	=	=	=
unicode_520	=	≠	=
japanese	=	≠	≠

だ、だれか
utf8mb4_japanese_ci を作って
(;´Д`)

おまけ

同じ文字とみなされるかどうかは weight_string() で確かめられる

utf8mb4_general_ci

```
mysql> select hex(weight_string('📚' collate utf8mb4_general_ci));
+-------------------------------------------------------+
| hex(weight_string('?' collate utf8mb4_general_ci)) |
+-------------------------------------------------------+
| FFFD                                                  |
+-------------------------------------------------------+

mysql>  select hex(weight_string('👍' collate utf8mb4_general_ci));
+-------------------------------------------------------+
| hex(weight_string('?' collate utf8mb4_general_ci)) |
+-------------------------------------------------------+
| FFFD                                                  |
+-------------------------------------------------------+
```

utf8mb4_unicode_520_ci

```
mysql> select hex(weight_string('📚' collate utf8mb4_unicode_520_ci));
+-------------------------------------------------------+
| hex(weight_string('?' collate utf8mb4_unicode_520_ci)) |
+-------------------------------------------------------+
| FBC3F363                                              |
+-------------------------------------------------------+

mysql> select hex(weight_string('📄' collate utf8mb4_unicode_520_ci));
+-------------------------------------------------------+
| hex(weight_string('?' collate utf8mb4_unicode_520_ci)) |
+-------------------------------------------------------+
| FBC3F37A                                              |
+-------------------------------------------------------+
```

おまけ２

パとﾊﾟ

- utf8_unicode_ci では「パ」＝「ハ゛」＝「ﾊ」
- 「パ」は一文字、「ﾊﾟ」は二文字
- 'ﾊﾟ' LIKE 'ﾊﾟ' => 偽
- 'ﾊﾟ' = 'ﾊﾟ' => **真**

= と LIKE は違うらしい

> *Per the SQL standard, LIKE performs matching on a per-character basis, thus it can produce results different from the = comparison operator*

http://dev.mysql.com/doc/refman/5.6/en/string-comparison-functions.html#operator_like

おわり