

メールアドレスのパーズ

2012/05/03

きちんとCcヘッダやToヘッダのメールアドレスを調べるには単純な `[^ ]+@[^ ]+` じゃだめなのでその辺りのメモ。正規表現のみでやろうとすると多分知らない苦労をする。

サンプルコード

\$Cc にある内容からメールアドレス、もしくはコメントを切り出す。結果は @addrs に格納される。Wikipediaの[メールアドレス](#)のページで利用可能な文字を加味した正規表現を作った。

[addrcheck.pl](#)

```
my $Cc =
'abc@hoge.jp","def@hoge.jp",jkl@hoge.jp,ThisIsTest,mno@hoge.jp,"this is
test"@hoge.jp';
my @addrs = $Cc =~ /"(?:\\\"|\\\\\\| [a-zA-Z0-9\\.!#$%&' *+|-|/=?\\^_`{|}~
\\(|)<>|[\\];:;,@])+"(?:!@)|\\((?:\\\"|\\\\\\| [a-zA-Z0-9\\.!#$%&' *+|-
|/=?\\^_`{|}~ \\(|)<>|[\\];:;,@])+"|[a-zA-Z0-9\\.!#$%&' *+|-
|/=?\\^_`{|}~|~|+@[a-zA-Z0-9\\. -]+|"(?:\\\"|\\\\\\| [a-zA-
Z0-9\\.!#$%&' *+|-|/=?\\^_`{|}~ \\(|)<>|[\\];:;,@])+"@[a-zA-Z0-9\\. -
]+/g;

$i=1;
print "****\\n$Cc\\n****\\n";
foreach my $ml ( @addrs )
{
    $a = ($ml =~ /^["\(\).+["\)]$/) ? " " : "*";
    printf("%03i: %1s [%s]\\n", $i, $a, $ml);
    $i++;
}
```

これを実行すると、

```
****
abc@hoge.jp,"def@hoge.jp",jkl@hoge.jp,ThisIsTest,mno@hoge.jp,"this is
test"@hoge.jp
****
001: * [abc@hoge.jp]
002:  ["def@hoge.jp"]
003: * [jkl@hoge.jp]
004: * [mno@hoge.jp]
005: * ["this is test"@hoge.jp]
```

となり、001,003,004,005 がメールアドレスと判断される。コメントまで検出しているのは、正規表現が面倒な事になるから。切り出したあとに、コメントを弾いてしまえばいいじゃない。

Wikipediaの[メールアドレス](#)のページによれば 005 の形式も正しい事になるらしい。

ここで切り出した他、ローカルパート部の検証

- ドットで始まっていないか
- ドットで終わっていないか
- ドットが連続していないか
- ローカルパート部の長さは64文字以内か

もしておくべきかもしれない。しかし、受信したメールのヘッダに記述のあるものなら、上記についてはクリアしていると考えても差し支えないと思う。

また“support@[192.0.2.69]”のようなドメイン部がブラケットで囲まれたIPアドレスもありらしいけど、実際に使われている様子が無いなら無視でもいいんじゃないかなと思う。

このサンプルコードでは実際にその形式は検出してない。

正規表現の説明

サンプルコードの正規表現では4つの文字列にマッチングする。メールアドレス(2形式)と、名前、コメント、の計4つ。

検出した内容から、メールアドレス形式ではないものを弾けばいい。無理にメールアドレスのみ検出しようとする、コメント内に書かれたメールアドレス的文字列をどのように排除すべきか、かなり悩む事になる。

ダブルクォートで囲まれ直後に『@』が無い文字列

ダブルクォートで囲ったコメント。直後に『@』があるとダブルクォートで囲まれたローカルパート部にもマッチングしてしまうため。

```
"(?:\\\"|\\\\\\\\| [a-zA-Z0-9\\.!#$%&'*+\\-\\/=\\?\\^\\_\\`\\\\{\\\\|\\\\}\\~ \\\\" data-bbox="72 579 921 596"/>
```

ダブルクォートの代わりに括弧『()』で囲まれた文字列

括弧で囲まれたコメントを検出。

```
\\((?:\\\"|\\\\\\\\| [a-zA-Z0-9\\.!#$%&'*+\\-\\/=\\?\\^\\_\\`\\\\{\\\\|\\\\}\\~ \\\\" data-bbox="72 702 887 719"/>
```

ローカルパート部がダブルクォートで囲まれていないメールアドレス

```
[a-zA-Z0-9\\.!#$%&'*+\\-\\/=\\?\\^\\_\\`\\\\{\\\\|\\\\}\\~]+@[a-zA-Z0-9\\.\\-]+
```

ローカルパート部がダブルクォートで囲まれているメールアドレス

```
"(?:\\\"|\\\\\\\\| [a-zA-Z0-9\\.!#$%&'*+\\-\\/=\\?\\^\\_\\`\\\\{\\\\|\\\\}\\~ \\\\" data-bbox="72 882 921 915"/>
```

[mail](#), [Perl](#), [技術資料](#)

From:

<https://wiki.hgotoh.jp/> - 努力したWiki

Permanent link:

<https://wiki.hgotoh.jp/documents/mail/mail-009>

Last update: **2023/04/14 02:32**

