

すみずみ君で業務効率が変わる！

正規表現検索のすすめ



インデックス作成不要で高速検索！ DFA型正規表現エンジン搭載
個人情報ファイル検出ツール **すみずみ君**



PC やサーバ内に日々増え続ける膨大なファイル、
必要なファイルをどのように探していますか？



Windows のフォルダ画面にある小さな**検索ボックスでファイルを検索**しています。

一般的に膨大なファイルの中から特定の文字列が含まれるファイルを探す場合、OS やアプリケーションの**全文検索機能**を使います。
検索時は **特定の文字列 = キーワード** を入力しているかと思えます。

例えば 電話番号 **03-1234-5678** を含むファイルを検索したい場合、**03-1234-5678** をキーワード検索すれば OK です。

では、**電話番号を含むファイルを見つけたい**場合、どうでしょうか。

電話番号は固定電話や携帯電話など膨大な数の番号があり、表記はいくつものパターンが存在します。

例 090-1234-5678 03(1234)5678 (045)1234-5678 08012345678

膨大な数と表記パターンが存在する電話番号を**キーワード検索**することは困難です。

電話番号のようにキーワードでは検索できない文字列を検索できるのが **正規表現検索** です。

正規表現 (せいきひょうげん : Regular Expression) とは

Regular = 規則的な } 検索したい文字列を
Expression = 表現 } という意味で **文字列のパターンで表現**することができます

実は正規表現はすでに 50 年以上もの歴史がある伝統的な検索方法なのです。

キーワード検索	正規表現検索	
信号機	信号機	キーワードと正規表現は同じ意味で、どちらも 信号機 を検索します。
信号機?	信号機?	キーワードと正規表現は違う意味で正規表現では 信号機 と 信号 を検索します。 これは ? がメタ文字で、 前の文字があってもなくても という表現です。

文字列のパターンで検索することができる大変便利な検索方法です

【携帯電話番号（携帯番号）を検索する正規表現】

1. 検索したい携帯番号の文字列パターン [090 - 1234 - 5678 ▶ 090 で始まるハイフン区切りの番号]

090 から始まるハイフン区切りの正規表現①

[0123456789] は角括弧 [] 内のどれか 1 文字にマッチする表現です。
下記の正規表現では 090-0000-0000 から 090-9999-9999 までの番号が検索できます。

090-[0123456789][0123456789][0123456789][0123456789]-[0123456789][0123456789][0123456789][0123456789]

090 から始まるハイフン区切りの正規表現②

[0123456789] は 0 から 9 までの連続数字なので [0-9] と表現できます。

090-[0-9][0-9][0-9][0-9]-[0-9][0-9][0-9][0-9] ※ [0-9] は半角数字 0 から 9 の 1 文字の表現です。

090 から始まるハイフン区切りの正規表現③

この正規表現では [0-9] が 4 回繰り返しているため [0-9]{4} と表現できます。

090-[0-9]{4}-[0-9]{4} ※ 波括弧 {} は 1 文字前の繰り返しの表現で、{4} は 1 文字前を 4 回繰り返す表現です。

090 から始まるハイフン区切りの正規表現④

半角数字だけの検索であれば [0-9] は %d と表現できます。

090-%d{4}-%d{4} ※ %d は 0 から 9 の半角数字 1 文字の表現です。

090、080、070 で始まるハイフン区切りの正規表現

携帯番号は 090 の他に 080 や 070 から始まる番号もありますので、080 と 070 の番号も検索する表現を追加します。

0[789]0-%d{4}-%d{4} ※ [789] は半角数字 7、8、9 のどれか 1 文字の表現です。

090、080、070 で始まるハイフン区切りおよびハイフンなし 11 桁の正規表現

携帯番号はハイフンなしのデータも存在しますので、ハイフンなしの番号も検索する表現を追加します。

0[789]0-?%d{4}-?%d{4} ※ -? は ? の前文字のハイフンがあってもなくても検索すると言う表現です。

この正規表現で
090、080、070 から始まる
半角数字ハイフン区切りとハイフンなし
11 桁の携帯番号を
検索することができます。



2. 検索したい携帯番号の文字列パターン [090(1234)5678 ▶ 半角括弧区切りの番号]

0[789]0%(%d{4}%)%d{4} ※ メタ文字の丸括弧 () 自身の表現には前にエスケープ文字の % を付けて %(と %) で表現します。

3. 検索したい携帯番号の文字列パターン [(090)1234-5678 ▶ 半角括弧区切りと半角ハイフン区切りの番号]

%([0789]0%)%d{4}-%d{4}

4. 検索したい携帯番号の文字列パターン [09012345678 ▶ 半角区切りなし 11 桁の番号]

0[789]0%d{8}

キーワードでは検索できない、電話番号、郵便番号、車両番号、製造番号、会員番号、文書番号など、
一定の規則で並んでいる文字列は正規表現で検索することができます。

【正規表現は数字列だけではなく、日本語文字列の検索も可能です】

例えば、下記 30 人の佐藤さんの名前を検索したい場合、全員を OR 検索するのは大変ですよ！

佐藤一郎、佐藤二郎、佐藤三郎、佐藤四郎、佐藤五郎、佐藤六郎、佐藤七郎、佐藤八郎、佐藤九郎、佐藤十郎、佐藤十一郎、佐藤十二郎、佐藤十三郎、佐藤十四郎、佐藤十五郎、佐藤十六郎、佐藤十七郎、佐藤十八郎、佐藤十九郎、佐藤二十郎、佐藤二十一郎、佐藤二十二郎、佐藤二十三元、佐藤二十四郎、佐藤二十五郎、佐藤二十六郎、佐藤二十七郎、佐藤二十八郎、佐藤二十九郎、佐藤三十郎

30 人の佐藤●郎さんの名前を検索する正規表現

正規表現では 佐藤 [どれか 1 文字] 郎 の表現で検索できます。

佐藤 [一二三四五六七八九十太多次治史志司嗣卓琢鉄徹哲利俊喜芳善嘉] 郎

30 人の佐藤●郎さんの他に、30 人の鈴木●郎さん、30 人の田中●郎さんの名前を検索する正規表現

(佐藤|鈴木|田中)[一二三四五六七八九十太多次治史志司嗣卓琢鉄徹哲利俊喜芳善嘉] 郎

(佐藤|鈴木|田中) は 佐藤 または 鈴木 または 田中 のいずれかを検索すると言う表現です。

高速ファイル検索ソフトウェア「すみずみ君」ではユーザ辞書として、キーワード辞書の他に**正規表現辞書も作成することができます。**
皆さまも大変便利な正規表現検索を是非すみずみ君で体験してみてください！

お問い合わせ

 三菱スペース・ソフトウェア株式会社

〒105-6132 東京都港区浜松町2-4-1 世界貿易センタービル32階
URL: <http://www.mss.co.jp>

■製品に関するお問い合わせ

営業本部 第一営業部

TEL: 0467-47-9103 FAX: 0467-47-9117

E-Mail: sumizumikun-info@mss.co.jp

*「すみずみ君」は三菱スペース・ソフトウェア株式会社の登録商標です。

*記載されている社名、商品名は全て各社の商標または登録商標です。

*内容は予告なしに変更することがありますので予めご了承ください。

*記載されている内容の一部または全部を複製、改変、転載することはいかなる理由を問わず禁止します。

