PostgreSQLでの正規表現によるテキストデータ抽出方法

## 【前提】

当資料では以下を前提に説明する。

* SQLの基本的な構文（SELECT文やWHERE句、AND、OR、LIKE演算子）は利用できる程度のレベルを想定
* 正規表現制約（先行/後方検索制約（(?=re)など））や正規表現量指定子（最短マッチ（\*?など））を極力利用しない
* プロシージャは利用しない

## 【基本】

### 1.SQLで正規表現を用いる構文

PostgreSQLで正規表現を用いた構文は以下の通りで、詳細は後述する。

1. パターンマッチメタ文字
2. 正規表現マッチ演算子
3. regexp\_match関数

（参考）SIMILAR TOによる正規表現パターンマッチ

標準SQLの正規表現定義を使用してパターンマッチする方法としてSIMILAR TOが存在するが、拡張機能に該当するPOSIX正規表現を適用できる正規表現マッチ演算子で説明するため、これについては記述しない。

#### 1.1.POSIX正規表現とは

Unixのsedやawkと類似したパターンマッチ言語を使用した正規表現。

#### 1.2.パターンマッチメタ文字

|  |  |
| --- | --- |
| メタ文字 | 説明 |
| | | 二者択一（2つの選択肢のうちいずれか） |
| \* | 直前の項目の0回以上の繰り返し |
| + | 直前の項目の1回以上の繰り返し |
| ? | 直前の項目の0回もしくは1回の繰り返し |
| {m} | 直前の項目の正確なm回の繰り返し |
| {m,} | 直前の項目のm回以上の繰り返し |
| {m,n} | 直前の項目のm回以上かつn回以下の繰り返し |
| 丸括弧() | 項目を1つの論理項目にグループ化 |
| 大括弧式[...] | POSIX正規表現と同様に文字クラスを指定 |

#### 1.3.正規表現クラス省略エスケープ

|  |  |
| --- | --- |
| エスケープ | 説明 |
| \d | 数字（全角を含む） |
| \D | 数字（全角を含む）以外 |
| \s | スペース（全角を含む） |
| \S | スペース（全角を含む）以外 |
| \w | 文字と数字それにアンダースコア |
| \W | 文字と数字それにアンダースコア以外 |

それぞれを否定する場合、大文字にすればよい。

#### 1.4.正規表現マッチ演算子

|  |  |  |  |
| --- | --- | --- | --- |
| 演算子 | 説明 | 例（True条件） | 例（False条件） |
| ~ | 正規表現にマッチ、大文字小文字の区別あり | 'abc' ~ 'a' | 'abc' ~ 'A' |
| ~\* | 正規表現にマッチ、大文字小文字の区別なし | 'abc' ~\* 'A' | 'abc' ~\* 'd' |
| !~ | 正規表現にマッチしない、大文字小文字の区別あり | 'abc' !~ 'd' | 'abc' !~ 'a' |
| !~\* | 正規表現にマッチしない、大文字小文字の区別なし | 'abc' !~\* 'd' | 'abc' !~\* ‘A' |

正規表現マッチ演算子の式は「[文字列] [演算子] [正規表現]」で記述する。

正規表現マッチ演算子の基本は「~」であり、大文字小文字を区別したくない場合は「\*」を付与し、それを否定する場合は「!」を付与すればよい。

#### 1.5.regexp\_match関数

|  |  |
| --- | --- |
| 構文 | regexp\_match([文字列], [正規表現], [フラグ])  ※[フラグ]は任意。詳細は「ARE埋め込みオプション文字」を確認。 |

regexp\_match関数はPOSIX正規表現パターンを文字列にマッチさせた結果、捕捉された最初の部分文字列のテキスト配列を返却する。フラグは「i」で大文字小文字を区別しないということは理解しておくこと。

（参考）regexp\_matches関数との違い

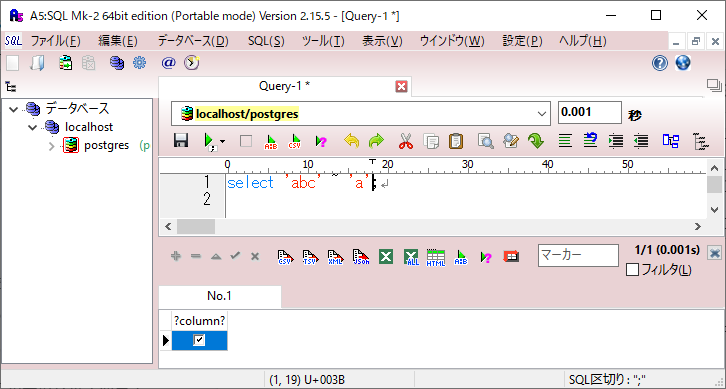
フラグに「g」を指定することで全ての部分文字列のテキスト配列を返却する点が異なる。

### 2.テストの仕方

select文でテストしたい正規表現を記述し、テストデータとして文字列を順次記述して確認を行う。

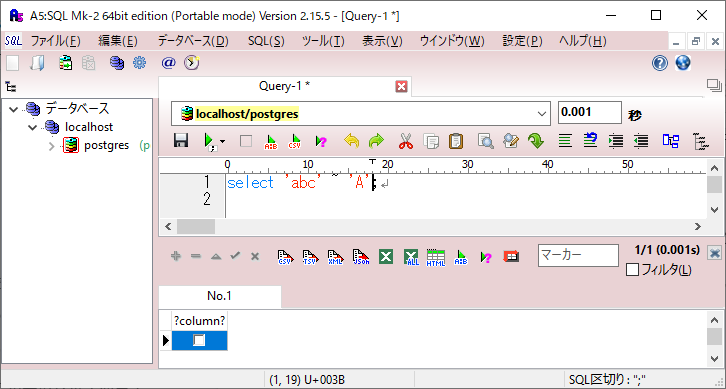
（例）「A」という文字列を大文字小文字区別せず検索する正規表現のテストの仕方

①「'abc' ~ 'a'」で正しく検索されることを確認する。

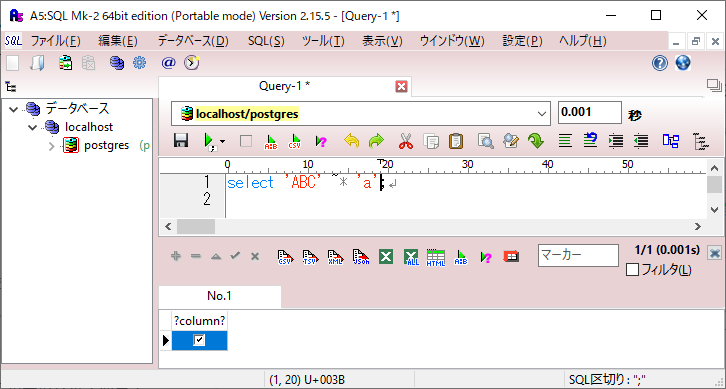


※A5SQLでTrueは上記のようにチェックボックスにチェックありで表記される。

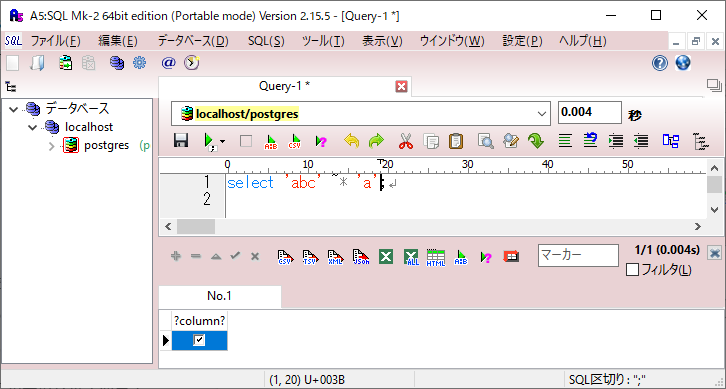
②「'ABC' ~ 'a'」だと検索されないため、正規表現が誤っていることを認識する。



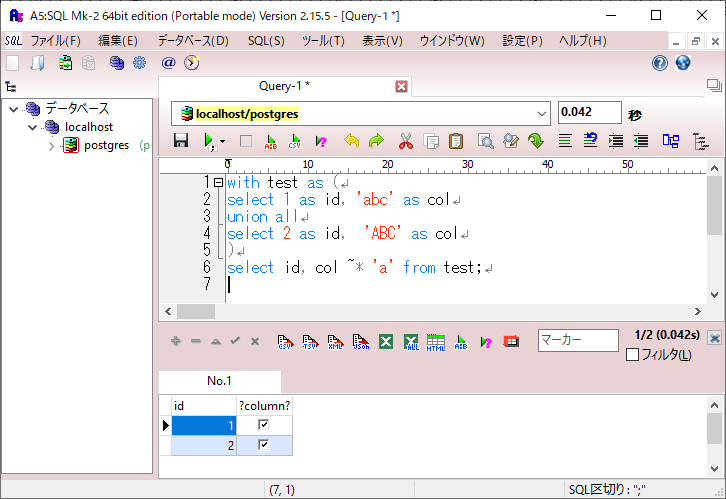
③「'ABC' ~\* 'a'」で正しく検索されることを確認する。



④「'abc' ~\* 'a'」でも正しく検索されることを確認する。



①´テストデータをまとめてWITH句で作成して実施することもできる。



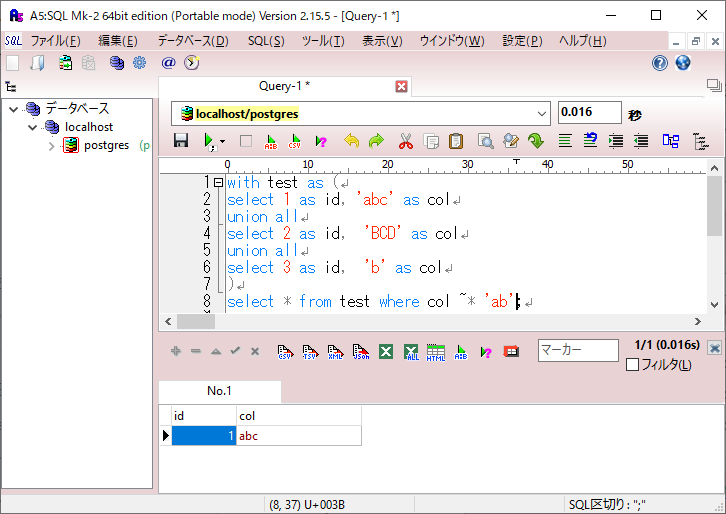
※「id」のカラムは結果を見やすくしているだけのため、記載しなくても問題ない。

## 【ユースケース】

### 1.対象レコード検索（基本）

#### 1.1.単一条件検索

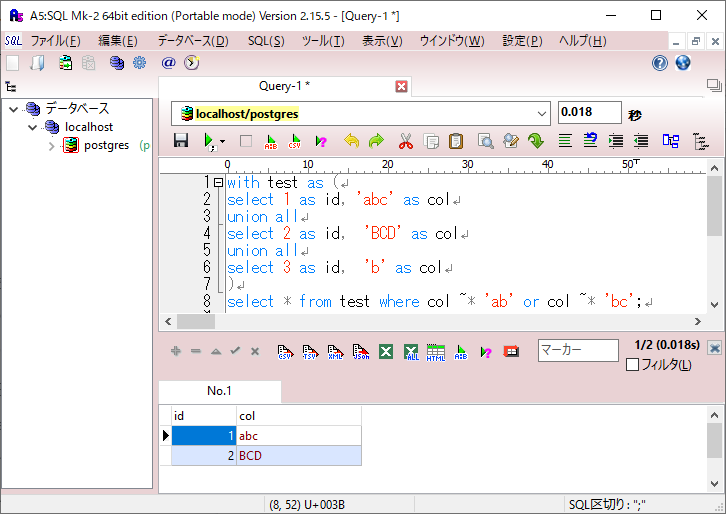
WHERE句で正規表現マッチ演算子を用いて、検索を行う。



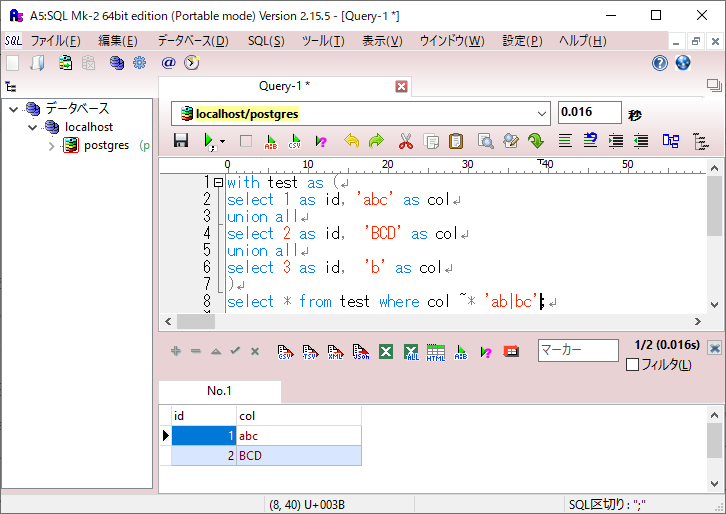
※id:1が正しく検索されている。

#### 1.2.複数条件検索（項目追加）

①OR条件で記載する。

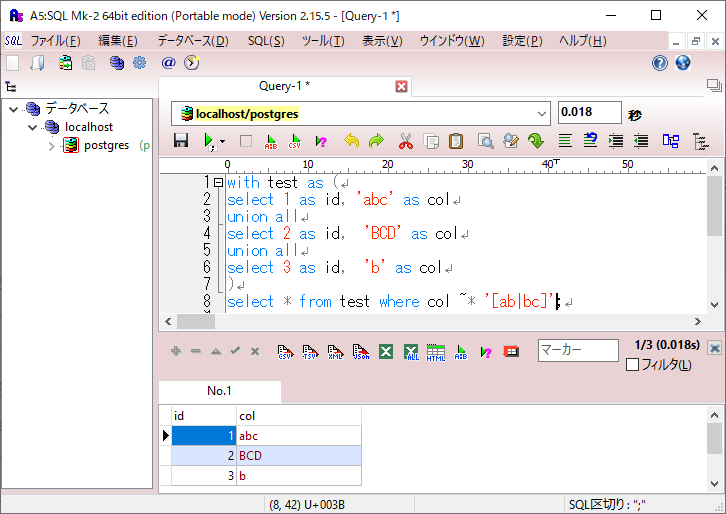


②メタ文字「|」を用いる。



（注意）メタ文字「大括弧式[]」

メタ文字の大括弧式は1文字の複数パターンをまとめたものなので注意が必要。

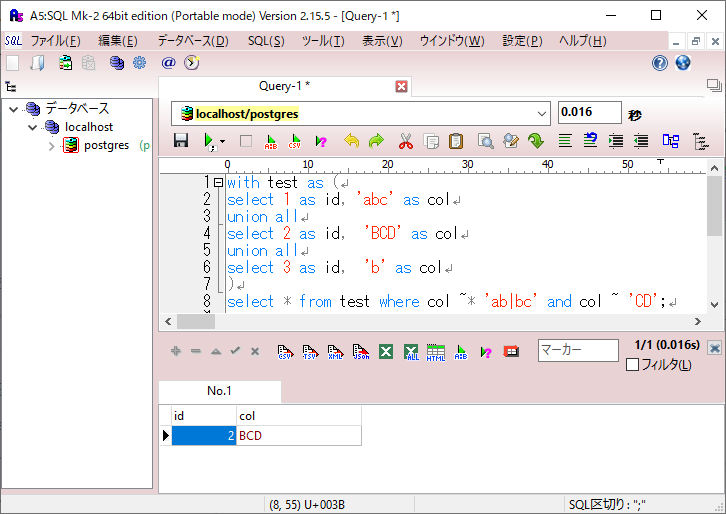


※「ab」または「bc」で検索するつもりが「b」も抽出されてしまっている。

　上記正規表現は「a」「b」「c」「|」を検索する正規表現となっている。

#### 1.3.複数条件検索（対象除外）

AND条件で否定を利用して除外する。



※大文字の「CD」だけは除きたい場合、AND条件で「CD」の否定形で除外している。

（注意）

④前後X文字

2.抽出

## 【参照元一覧】

* PostgreSQL 10.5文書 第9章 関数と演算子 9.7. パターンマッチ  
  <https://www.postgresql.jp/docs/10/functions-matching.html>
* PostgreSQLで正規表現  
  <https://www.postgresql.jp/docs/10/functions-matching.html>