**言語データ分析の方法**

**検索・置換・集計・行列・確率**

ver. 2015.4.16

この文書は「言語データ分析」の１学期間のテキストをまとめたものです。授業の進行に従って随時改訂していきますので、常に最終バージョンを参照してください。ご使用になられた方はぜひご意見をお寄せください。私のメールアドレスはLETRASの開始ページをご覧ください。参考にさせていただき、よりよいものを目指したいと思います。よろしくお願いいたします。

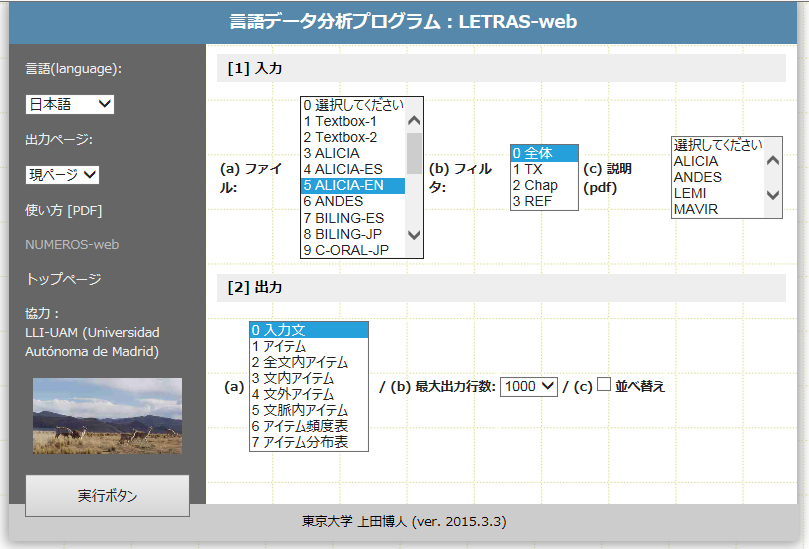
http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/

(東京大学・上田博人)

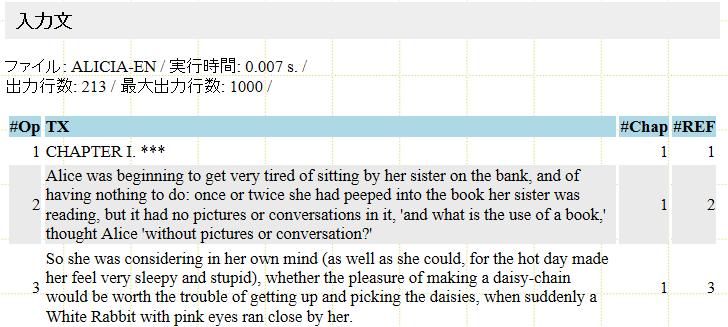
# 簡単な検索

## 入力文

はじめに次のサイトを使って簡単な検索の練習からはじめます。



左上の**言語**から「日本語」を選択し、**[1]入力**の**(a)ファイル**から「5 ALICIA-EN」(Alice's Adventures in Wonderland, Lewis Carroll, 1895の抜粋)を選択してください。そのまま実行ボタンを押すと次のように文書が出力されます。



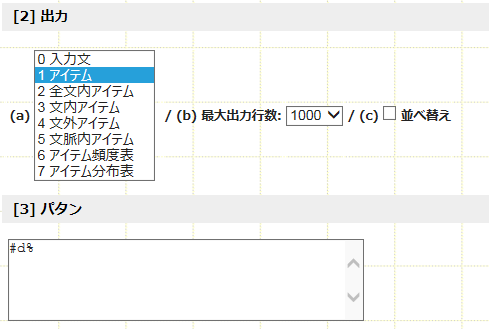
以下では、この文書を使って簡単な検索の練習をします。上のようなテキストデータの出力を、次のような表形式にして説明します。

|  |  |  |  |
| --- | --- | --- | --- |
| #Op | TX | #Chap | #REF |
| 1 | CHAPTER I. \*\*\* | 1 | 1 |
| 2 | Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?' | 1 | 2 |
| 3 | So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her. | 1 | 3 |

最初の行を「タイトル行」と呼びます。タイトル行の#Opは出力行、TXはテキスト、Chapは原作の章、REFは入力データの参照番号を指します。これは出力の形式が「入力文」なので、入力データそのものが出力されます。よって、出力行(#Op)(と参照番号(#REF)は同じになります。

## アイテム・連立検索式・出力の並べ替え

2.出力で「アイテム」を選択してください。検索された文字列を「アイテム」とよびます。そうすると次のように[3]パタンが出現します。



ここに「#d%」が書かれていますが、これを消して、Aliceと書き込み、実行ボタンをクリックすると、次が出力されます。

|  |  |  |  |
| --- | --- | --- | --- |
| #Op | アイテム | #Chap | #REF |
| 1 | Alice | 1 | 2 |
| 2 | Alice | 1 | 2 |
| 3 | Alice | 1 | 4 |
| 4 | Alice | 1 | 4 |
| 5 | Alice | 1 | 5 |

この表からAliceが出現する章と参照番号がわかります。

【練習-1】英語の高頻度語、the, and, ofなどを検索しましょう。

【練習-2】Aliceとrabbitを一緒に検索しましょう。2つ以上の検索をするときのパタンは、次のように改行してください。これを**連立検索式**とよびます。

Alice

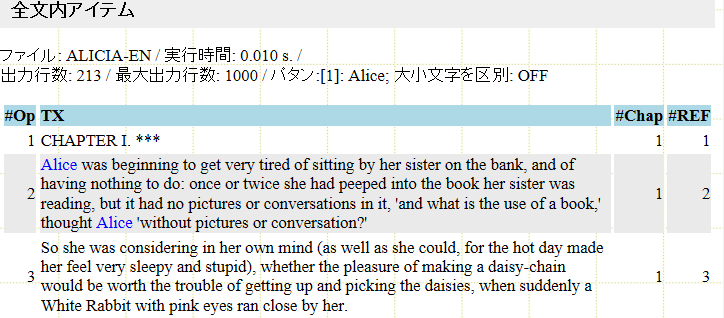
rabbit

【練習-3】**[2]出力の並べ替え**にチェックを入れて、出力を比較しましょう。このように、出力には**入力順**と**並べ替え**が選択できます。それぞれのどのような使い方があるか考えてみましょう。

## 文とアイテム・大小文字の区別

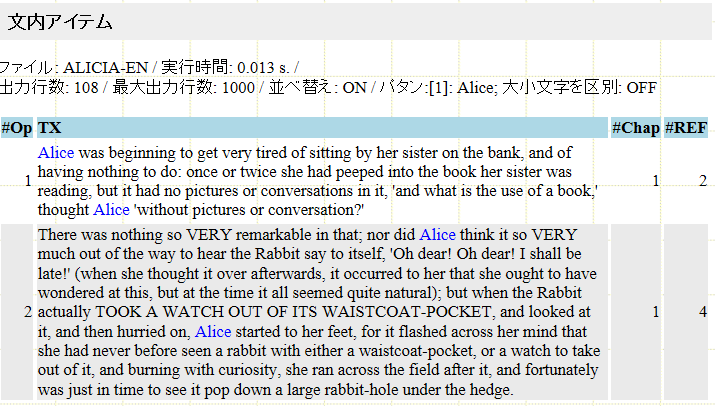
### 全文内アイテム

[2]出力から全文内アイテムを選択し、Aliceを検索します。実行すると次を出力します。全文内アイテムでは、アイテムの有無にかかわらず、文全体を出力し、該当部分をマークします。次の出力では、#REF=3ではAliceは出現していません。



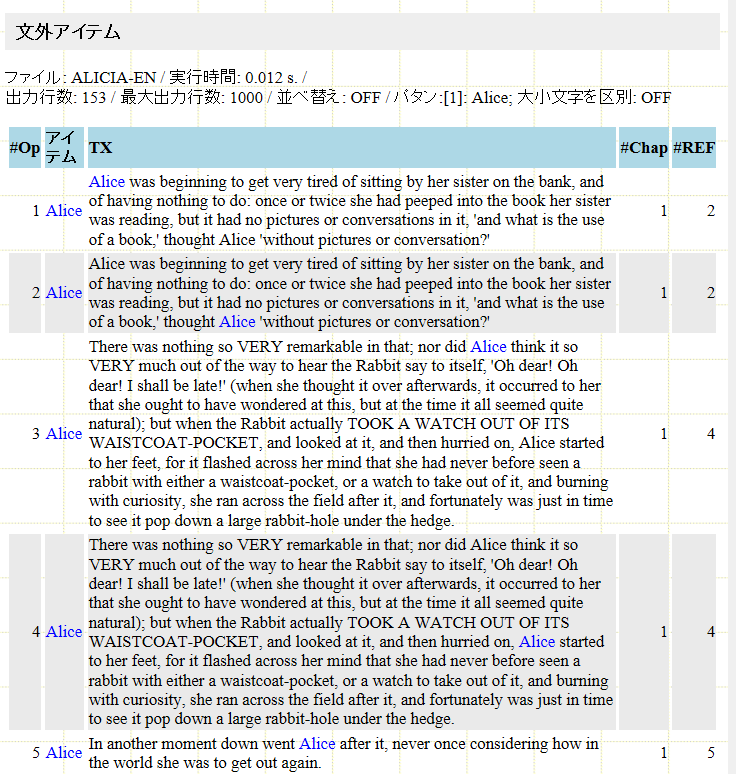
### 文内アイテム

一方、**文内アイテム**を選択すると、次のようにアイテムがある文だけを出力します。先に見たように、#REF=3ではAliceが出現していないので、#REFは2, 4のように非連続になります。



### 文外アイテム

出力に文外アイテムを選択すると、アイテムと文を並置して出力します。複数のアイテムが出現する文は、それぞれ個別に出力されます。



【練習-1】**文外アイテム**を使って、**[3] パタン**の**大小文字の区別**をチェックしRabbitとrabbitの出力を比べなさい。次の4つの場合を確かめましょう。

(1) パタン：Rabbit + 大小文字の区別なし

(2) パタン：Rabbit + 大小文字の区別あり

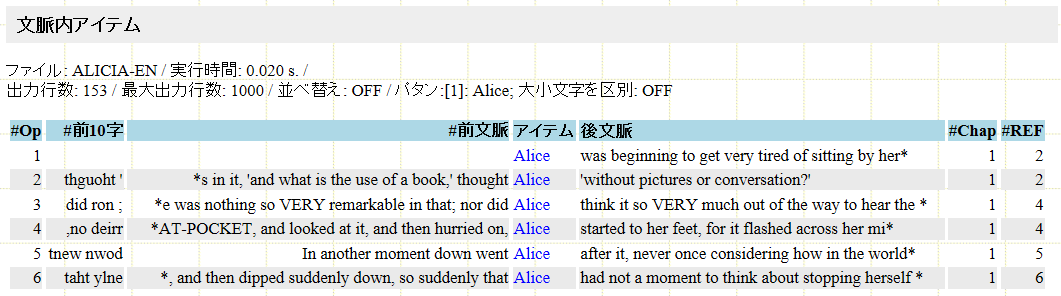
(3) パタン：rabbit + 大小文字の区別なし

(4) パタン：rabbit + 大小文字の区別あり

【練習-2】全文内アイテム・文内アイテム・文外アイテムの使い方について考察しなさい。

## 文脈内アイテム・列の並べ替え

出力に文脈内アイテムを選択すると、アイテムが前文脈と後文脈に挟まれて出力されます。アイテムが使われる文脈を探るのに有効です。



　前文脈と後文脈は一定の長さでカットされます。

　上の出力のタイトル行（青の背景色）のセルをクリックすると、昇順(ユニコード順：一般のABC順)・降順（逆順）で並べ替えをします。

　「前10字」は「前文脈」の最後の10文字を逆に配列したもので、これを並べ替えると「逆引き配列」になります。

|  |  |  |  |
| --- | --- | --- | --- |
| #Op | #前10字 | #前文脈 | アイテム |
| 1 |  |  | Alice |
| 12 |  |  | Alice |
| 18 | '!gnitiaw | \*h! won't she be savage if I've kept her waiting!' | Alice |
| 59 | '!puos ta | There's certainly too much pepper in that soup!' | Alice |
| 8 | ,ees uoy | \*our thousand miles down, I think—' (for, you see, | Alice |
| 140 | ,esruoc f | \*urors had a pencil that squeaked. This of course, | Alice |
| 86 | .der meht | \* three gardeners at it, busily painting them red. | Alice |
| 122 | .elbmert | \*e arm that was linked into hers began to tremble. | Alice |
| 32 | deciton t | Very soon the Rabbit noticed | Alice |
| 142 | deirc '!e | Here!' cried | Alice |
| 14 | deirc '!r | Curiouser and curiouser!' cried | Alice |
| 116 | demialcxe | Oh, I know!' exclaimed | Alice |
| 101 | detuohs ' | \*ion was evidently meant for her. / 'Yes!' shouted | Alice |
| 99 | dias '!de | \* for protection. / 'You shan't be beheaded!' said | Alice |
| 102 | dna ,neeu | Come on, then!' roared the Queen, and | Alice |
| 89 | dna ,spet | \*r faces. There was a sound of many footsteps, and | Alice |
| 74 | elihw ,et | \*ped, and the party sat silent for a minute, while | Alice |
| 11 | ereh dnA | \*u know. But do cats eat bats, I wonder?' And here | Alice |
| 40 | hsiloof u | Oh, you foolish | Alice |
| 90 | meht gnom | \*e guests, mostly Kings and Queens, and among them | Alice |
| 87 | nopu llaf | \*unjust things—' when his eye chanced to fall upon | Alice |
| 108 | nopu nihc | \*as exactly the right height to rest her chin upon | Alice |
| 10 | os ,od ot | \*own, down, down. There was nothing else to do, so | Alice |
| 121 | ot ,ereh | \* 'as pigs have to fly; and the m—' / But here, to | Alice |
| 16 | rooP | Poor | Alice |

上の出力は、上下のデータを適当に抽出したものです。これを見るとはじめに前文脈がない行があり、次にシングルコーテーション、コンマ、ピリオドが続き、その下に末尾がd, e, h, m, n, o, rの順で出力されていることがわかります。また同じ末尾のデータ間では、たとえば、末尾がdのデータは、noticed, cried, exclaimed, shoutedの順になります。これらは末尾から2番目の文字、3番目の文字、…というように比較し、それらをABC順に並べられています。この配列はアイテムの前文脈にどのような文字・記号・語列があるのかを調べるときに役立ちます。

【練習-1】act, fact, forget, get, patをWordだけと使って逆引き配列に並べ替えなさい。

【練習-2】アイテムにcried, said, thoughtを指定し、出力を「前10字」・「アイテム」・「後文脈」で並べ替えて出力データを分析しなさい。

## 集計

# 正規表現による検索

正規表現(regular expression)を使うことによって簡単な検索（文字列による検索）ではできないことが可能になります。正規表現の規則は少なく非常に単純ですが使い方については練習が必要です。何度でも実験して確認してください。

## 正規表現

### 特殊文字

|  |  |
| --- | --- |
| **\b** | 任意の英単語の境界に一致します。 |
| **\B** | 任意の英単語境界以外の位置に一致します。 |
| **\n** | 改行に一致します。 |

**入力文：**

|  |
| --- |
| The Universal Declaration of Human Rights Article 1. All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood. |

正規表現：\bin\b：単語境界に挟まれたin

|  |
| --- |
| The Universal Declaration of Human Rights Article 1. All human beings are born free and equal {\***in\*}** dignity and rights. They are endowed with reason and conscience and should act towards one another {\***in\*}** a spirit of brotherhood. |

beingsの中のinは、単語境界に挟まれていないので一致しません。

### 文字クラス

|  |  |
| --- | --- |
| **[*xyz*]** | 文字セットに含まれている任意の1文字に一致します。[...]の中では., ? , \*などをエスケープする必要はありません。[a-d]は[abcd]と同じです。 |
| **[^*xyz*]** | 文字セットに含まれていない任意の1文字に一致します。 |
| **.（ピリオド）** | 改行(\n)以外の任意の文字に一致します。 |
| **\w** | 英単語に使用される任意の文字（アルファベット、数字、アンダースコア）[a-zA-Z0-9\_]に一致します。 |
| **\W** | 英単語に使用される文字以外の任意の文字に一致します。  [^a-zA-Z0-9[a-zA-Z\_0-9]と同じです。 |
| **\d** | 任意の数字に一致します。[0-9]と同じです。 |
| **\D** | 任意の数字以外の文字に一致します。[^0-9]と同じです。 |
| **\s** | 任意のスペース文字に一致します。[ \t\r\n\f]と同じです。 |
| **\S** | 任意の非スペース文字に一致します。[^ \t\r\n\f]と同じです。 |

**入力文：**

|  |
| --- |
| All human beings are born free and equal in dignity and rights. |

正規表現検索：[e-h] （e, f, g, h, iに一致）

|  |
| --- |
| All {\*h\*}uman b{\*e\*}in{\*g\*}s ar{\*e\*} born {\*f\*}r{\*e\*}{\*e\*} and {\*e\*}qual in di{\*g\*}nity and ri{\*g\*}{\*h\*}ts. |

正規表現検索：[＾a-v] （a-v以外に一致）

|  |
| --- |
| All{\* \*}human{\* \*}beings{\* \*}are{\* \*}born{\* \*}free{\* \*}and{\* \*}equal{\* \*}in{\* \*}dignit{\*y\*}{\* \*}and{\* \*}rights{\*.\*} |

正規表現検索：[c-i] （c, d, e, f, g, h, iに一致）

|  |
| --- |
| All human beings are born free and equal in dignity and rights. |

正規表現検索：[^c-i]（c, d, e, f, g, h, i以外に一致）

|  |
| --- |
| All human beings are born free and equal in dignity and rights. |

### 選択、グループ化、繰り返し

|  |  |
| --- | --- |
| **|** | 複数の句を1つの正規表現にまとめ、これらのうちの任意の句に一致します。たとえば、d(os|a)はdosまたは daに一致します。\ba\b|\bthe\bのように(…)でも使うことができます。 |
| **(…)** | 複数の句をグループ化して1つの句を作成します。(ab)\*cは abc またはc に一致します。 |
| **+** | 1個以上の直前の文字に一致します。{1,}と同じです。e+でe, ee, eee, …に一致します。 |
| **\*** | ゼロ個以上の直前の文字またはぐるに一致します。{0,}と同じです。ah\*でa, ah, ahh, …に一致します。 |
| **?** | ゼロ個または1個の直前の文字に一致します。{0,1}と同じです。books?でbookとbooksに一致します。 |
| **{*a*}** | 先行する正規表現*a*個に一致します。[aeoiu]{2}で2母音の連続（ei, ee, uaなど）に一致します。 |
| **{*a*,}** | 先行する正規表現*a*個以上の直前の文字に一致します。[aeoiu]{3,}で3母音の連続（aei, uai, auuuなど）に一致します。 |
| **{*a,b*}** | 先行する正規表現*a*個以上、*b*個以下に一致します。[aeoiu]{2, 4}で2-4母音の連続（ei, aei, uai, auuuなど）に一致します。 |

正規表現検索：(free|equal) （freeとequalに一致します。）

|  |
| --- |
| All human beings are born {\*free\*} and {\*equal\*} in dignity and rights. |

正規表現検索：[e-h]+　（[e-h]の連続に一致します。）

|  |
| --- |
| All {\*h\*}uman b{\*e\*}in{\*g\*}s ar{\*e\*} born {\*f\*}r{\*ee\*} and {\*e\*}qual in di{\*g\*}nity and ri{\*gh\*}ts. |

正規表現検索：[aeoiu]{2} （2母音の連続に一致します。）

|  |
| --- |
| All human b{\*ei\*}ngs are born fr{\*ee\*} and eq{\*ua\*}l in dignity and rights. |

正規表現検索：[^aeoiu]{2,}　（母音以外の文字2個以上の連続に一致）

|  |
| --- |
| A{\*ll h\*}uma{\*n b\*}ei{\*ngs \*}are{\* b\*}o{\*rn fr\*}ee a{\*nd \*}equa{\*l \*}i{\*n d\*}i{\*gn\*}i{\*ty \*}a{\*nd r\*}i{\*ghts.\*} |

### エスケープ文字

特殊文字の検索(, ), [, ], {, }, ?, !, .(ピリオド), +, \*, |, \を探すときは，その前に\をつけてエスケープします。たとえば\?でクエスチョンマークを検索します。

入力文：

|  |
| --- |
| ¿Cómo está usted? |

正規表現：\?（クエスチョンマーク）

|  |
| --- |
| ¿Cómo está usted{\*?\*} |

### 置換文字

正規表現の後方参照を使うと検索式の一部を参照することができます。参照する正規表現を丸括弧で(…)囲み、$の後に数字を続けることによってその句を指定します。

|  |  |
| --- | --- |
| **$*n*** | 検索パタンの*n*番目の(…)に一致した文字列 |
| **$$** | $という文字 |

入力文：

|  |
| --- |
| Rumi: Hola, profesor.Prof. Rubio: Buenos días.Rumi: Buenos días. Nos encontramos ahora en la Universidad Complutense de Madrid. ¿Dónde nos vamos ahora?  Prof.: Bueno, vamos a iniciar hoy el Camino del Cid, la primera parte. |

正規表現：#(c\w)=>[[$1]]（cで始まる語を[[...]]でマークします。）

|  |
| --- |
| Rumi: Hola, profesor.Prof. Rubio: Buenos días.Rumi: Buenos días. Nos encontramos ahora en la Universidad [[Complutense]] de Madrid. ¿Dónde nos vamos ahora?  Prof.: Bueno, vamos a iniciar hoy el [[Camino]] del [[Cid]], la primera parte. |

次は中世スペイン語の資料を文字化した資料です。

|  |
| --- |
| Otro(22)ssí mando que los menestrales non echen suerte en el judgado por seer juezes, ca el juez deve tener la seña, e tengo que si <a> afruenta viniesse o a logar de periglo e omne vil o rafez toviesse la seña que podrié (23) caer el concejo en grant onta e en grant vergüença. |

冒頭のOtro(22)ssíは語の途中で改行されssí以下が22行目になることを示しています。文法研究のためには、これをOtrossí (22)にする必要があります。これは次の置換式によって実現できます。

正規表現：(\(\d+\))(&)=>$2 $1

|  |
| --- |
| Otrossí (22) mando que los menestrales non echen suerte en el judgado por seer juezes, ca el juez deve tener la seña, e tengo que si <a> afruenta viniesse o a logar de periglo e omne vil o rafez toviesse la seña que podrié (23) caer el concejo en grant onta e en grant vergüença. |

### 後方参照

後方参照を使うことで式の内容を記憶させ、それを後から参照させることができます。

|  |  |
| --- | --- |
| (...)\*n* | 検索文字列の(...)の式に一致した文字列が記憶され、それを*n*回繰り返して参照します。 |
| (...)...(...)=>$*n* | 検索文字列の(...)の式に一致した文字列が記憶され、置換文字列でそれを参照して再生します。*n*は(...)の順番に対応する番号です。 |

入力文 ：

|  |
| --- |
| どんどんテーマが広がって、ますます興味がわいてきた。 |

正規表現：(..)\1：2文字が2回繰り返す文字列

|  |
| --- |
| どんどんテーマが広がって、ますます興味がわいてきた。 |

参照する文字（列）がわかっているときは、検索式をたとえば「(どん){2}」のようにすることができますが、ここでは他にも「ますます」「ぐんぐん」のように、さまざまに変化する場合を想定しています。\1が先行する(..)を後方から参照しています。

## 拡張正規表現

### 特殊文字を再定義

LETRAS.xlsmでは一般の正規表現を拡張して次の検索字を再定義します。

|  |  |
| --- | --- |
| **#** | 単語の境界：#b%はbで始まる単語を検索します。 |
| **##** | 文の境界：\aを区切りとします。 |
| **&** | 単語文字1個以上 \l+ |
| **%** | 単語文字0個以上 \l\* |
| **\l** | 西欧語単語文字 [A-Za-zÀ-ǿ] |
| **\L** | 西欧語単語文字以外 [^A-Za-zÀ-ǿ] |
| **\i** | キリル文字 |
| **\I** | キリル文字以外 |
| **\g** | ギリシャ文字 |
| **\G** | ギリシャ文字以外 |
| **\e** | ハングル |
| **\E** | ハングル以外 |
| **\y** | CJK互換漢字、統合漢字、漢字拡張文字 |
| **\Y** | CJK互換漢字、統合漢字、漢字拡張文字以外 |
| **\v** | 母音文字 [aeiouÀ-Æà-æÈ-Ëè-ëÌ-Ïì-ïÒ-Öò-öÙ-Üù-ü] |
| **\V** | 母音文字以外 [^aeiouÀ-Æà-æÈ-Ëè-ëÌ-Ïì-ïÒ-Öò-öÙ-Üù-ü] |
| **\c** | 子音文字 [bcdfghj-np-tv-zÇçÑñß] |
| **\C** | 子音文字以外 [^bcdfghj-np-tv-zÇçÑñß] |
| **\a** | 文頭句読点：[¿¡] |
| **\A** | 文頭句読点以外：[^¿¡] |
| **\p** | 句読点：[.:?!。．？！：] |
| **\P** | 句読点以外：[^.:?!。．？！：] |
| ***n*@** | n字連続 |
| ***n*&** | n語連続 |

### 単語の境界

拡張正規表現：#(m%)=>[$1]

|  |
| --- |
| Otro(22)ssí [mando] que los [menestrales] non echen suerte en el judgado por seer juezes, ca el juez deve tener la seña, e tengo que si <a> afruenta viniesse o a logar de periglo e omne vil o rafez toviesse la seña que podrié (23) caer el concejo en grant onta e en grant vergüença. |

拡張正規表現：%[rs]

|  |
| --- |
| Otro(22)ssí [mando] que {\*los\*} [{\*menestrales\*}] non echen suerte en el judgado {\*por\*} {\*seer\*} {\*juezes\*}, ca el juez deve {\*tener\*} la seña, e tengo que si <a> afruenta viniesse o a {\*logar\*} de periglo e omne vil o rafez toviesse la seña que podrié (23) {\*caer\*} el concejo en grant onta e en grant vergüença. |

### 日本語文字の再定義

日本語文字を検索するときは、\h（ひらがな）, \k（カタカナ）, \z（漢字）, \j（日本語文字）を使用します。

|  |  |
| --- | --- |
| **\h** | ひらがな [ぁ-んー] |
| **\H** | ひらがな以外 [^ぁ-んー] |
| **\k** | カタカナ [ァ-ンー] |
| **\K** | カタカナ以外 [^ァ-ンー] |
| **\z** | 漢字 [一-龥々〆] |
| **\Z** | 漢字以外 [＾一-龥々〆] |
| **\j** | 日本語文字 [ぁ-んァ-ンー一-龥々〆] |
| **\J** | 日本語文字以外 [^ぁ-んァ-ンー一-龥々〆] |

入力文：『坊っちゃん』夏目漱石

|  |
| --- |
| 親譲りの無鉄砲で小供の時から損ばかりしている。 |

拡張正規表現 \z{2}（漢字2文字の連続）

|  |
| --- |
| {\*親譲\*}りの{\*無鉄\*}砲で{\*小供\*}の時から損ばかりしている。 |

拡張正規表現:\z+（1個以上の漢字）

|  |
| --- |
| {\*親譲\*}りの{\*無鉄砲\*}で{\*小供\*}の{\*時\*}から{\*損\*}ばかりしている。 |

拡張正規表現 \z\*\h+（漢字とひらがな：±漢字＋ひらがな）

|  |
| --- |
| 『{\*坊っちゃん\*}』夏目漱石  {\*親譲りの\*}{\*無鉄砲で\*}{\*小供の\*}{\*時から\*}{\*損ばかりしている\*}。 |

### 外国語文字の再定義

「再定義」というシートには次のような設定をしています。これは自由に変えることができます。変更したときは「更新」ボタンを押してください。ここで使用している「/, ", ~, `, ^」という特殊記号を検索するときは前に\をつけてエスケープしてください。

|  |  |
| --- | --- |
| \/ | x128$ |
| a/ | á |
| e/ | é |
| i/ | í |
| ó | ó |
| ú | ú |
| A/ | Á |
| E/ | É |
| I/ | Í |
| Ó | Ó |
| Ú | Ú |
| \" | x128$ |
| a" | ä |
| e" | ë |
| i" | ï |
| o" | ö |
| u" | ü |
| A" | Ä |
| E" | Ë |
| I" | Ï |
| O" | Ö |
| U" | Ü |
| x128$ | " |
| \~ | x128$ |
| a~ | ã |
| e~ | e |
| i~ | i |
| o~ | õ |
| u~ | u |
| A~ | Ã |
| E~ | E |
| I~ | I |
| O~ | Õ |
| U~ | U |
| x128$ | ~ |
| \` | x128$ |
| a` | à |
| e` | è |
| i` | ì |
| o` | ò |
| u` | ù |
| A` | À |
| E` | È |
| I` | Ì |
| O` | Ò |
| U` | Ù |
| x128$ | ` |
| \^ | x128$ |
| a^ | â |
| e^ | ê |
| i^ | î |
| o^ | ô |
| u^ | û |
| A^ | Â |
| E^ | Ê |
| I^ | Î |
| O^ | Ô |
| U^ | Û |
| x128$ | ^ |

### 前後の条件

|  |  |
| --- | --- |
| {{正規表現}} | 検索文字列の前後に付加して検索の条件とします。 |

たとえば、{{te }}va%はteと空白に続くデータを検索します。

### 連続検索

たとえば正規表現「@@@」（3文字）を使うと、recepciónの中でそれにマッチした字列を次のように次々に出力します。

|  |  |
| --- | --- |
| Sch.Flm. | **@@@** |
| @@@ | **rec** |
| @@@ | **epc** |
| @@@ | **ión** |

このとき、正規表現「@@@」にマッチした字列は、次の検索の対象になりません。そこで、recepciónはrec, epc,iónのように区切られて出力されます。

　一方、3@にように、最初に数字をつけて検索すると、次のような出力になります。

|  |  |
| --- | --- |
| Sch.Flm. | **3@** |
| 3@ | **rec** |
| 3@ | **ece** |
| 3@ | **cep** |
| 3@ | **epc** |
| 3@ | **pci** |
| 3@ | **ció** |
| 3@ | **ión** |

このときは、recepciónはrec, ece, cep, epc, pci, ció, iónのように、あらゆる3文字連続が出力されます。このような出力を「連続検索」とよぶことにします。連続検索の検索式は、字連続の場合は「3@」のように、数字＋@とし、語連続の場合は「3&」のように、数字＋&とします。連続検索は、アイテムの他に、以下の文外アイテム、文脈内アイテム、語列内アイテム、および「集計」で使うことができます。

### 語と文の長さ

拡張正規表現を使って語と文の長さを指定し、検索したり、集計したりすることができます。文の長さについては、長さの単位を文字数と語数で指定します。このとき、置換機能によって文の区切りを設定してから検索すると正確な検索・集計ができます。

#### (1) 語長（文字数）：〔集計〕絶対頻度

たとえば、「#@{2}#」という検索式の#は単語境界を示し、@は単語文字を示し、{...}で繰り返し数を示します。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| \* | Madrid | Sevilla | México | Lima | B.A. | 計 |
| #@{1}# | 354 | 338 | 420 | 254 | 302 | 1,668 |
| #@{2}# | 1,429 | 1,374 | 1,640 | 1,022 | 1,286 | 6,751 |
| #@{3}# | 836 | 847 | 975 | 624 | 799 | 4,081 |
| #@{4}# | 643 | 602 | 773 | 500 | 587 | 3,105 |
| #@{5}# | 712 | 694 | 855 | 612 | 713 | 3,586 |
| 計 | 3,974 | 3,855 | 4,663 | 3,012 | 3,687 | 19,191 |

#### (2) 文長（文字数）：〔集計〕絶対頻度

はじめに置換式「(.\*?\p +)=><$1>」によって、文末符号で終わる文を<…>で囲みます。「.\*?」で任意の文字列を示し、その後に文末符号(\p)と1個以上のブランクがある連続(\p +)を$1によって後方から参照し、その両側に<…>を付加します。2番目からは検索式になります。はじめの検索式は「<.{1,10}>」で、<…>で囲まれた文字の数が、1個以上、10個以下である文を検索します。以下も同様です。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| (.\*?\p +)=><$1> <.{1,10}> <.{21,30}> <.{31,40}> <.{41,50}> | Madrid | Sevilla | México | Lima | B.A. | 計 |
| <.{1,10}> | 18 | 16 | 39 | 18 | 25 | 116 |
| <.{21,30}> | 78 | 55 | 104 | 56 | 84 | 377 |
| <.{31,40}> | 60 | 56 | 86 | 63 | 59 | 324 |
| <.{41,50}> | 61 | 39 | 88 | 39 | 45 | 272 |
| 計 | 217 | 166 | 317 | 176 | 213 | 1,089 |

＊文データに<…>が使われているときには、«…»などで文を囲みます。

#### (3) 文長（語数）：〔集計〕絶対頻度

先と同様に、はじめに置換式「(.\*?\p +)=><$1>」によって、文末符号で終わる文を<…>で囲みます。次に、置換式「(&)=>{$1}」によって、単語(&)の両側を{…}で囲みます。

　3行目から以下は検索式です。たとえば「<\L\*({&}\L\*){2}\L\*>」は非単語文字列が0個以上あって(\L\*)、「単語と0個以上の非単語文字列」という文字連続が2個（({&}\L\*){2}）、その後に0個以上の非単語文字列(\L\*)、という検索式になります。これにより、次のような文字列が検索されます。

<– ¡{Buenos} {días}! >

<– {Bien}, {señor}. >

<{Muchas} {gracias}. >

次が1語文、2語文、3語文、4語文の集計です。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| (.\*?\p +)=><$1> (&)=>{$1} <\L\*({&}\L\*){1}\L\*> <\L\*({&}\L\*){2}\L\*> <\L\*({&}\L\*){3}\L\*> <\L\*({&}\L\*){4}\L\*> | Madrid | Sevilla | México | Lima | B.A. | 計 |
| <\L\*({&}\L\*){1}\L\*> | 14 | 16 | 41 | 22 | 34 | 127 |
| <\L\*({&}\L\*){2}\L\*> | 51 | 45 | 81 | 46 | 40 | 263 |
| <\L\*({&}\L\*){3}\L\*> | 51 | 30 | 80 | 34 | 40 | 235 |
| <\L\*({&}\L\*){4}\L\*> | 46 | 36 | 65 | 27 | 43 | 217 |
| 計 | 162 | 127 | 267 | 129 | 157 | 842 |