

全文検索システム『ひまわり』/利用者マニュアル/1_6/6. アノテーション内容を集計する

[Top](#) / [全文検索システム『ひまわり』](#) / [利用者マニュアル](#) / [1_6](#) / 6. アノテーション内容を集計する

言語を選択 ▼

[Prev](#)

[全文検索システム『ひまわり』/利用者マニュアル/1_6](#)

[Next](#)

6. アノテーション内容を集計する [↑]

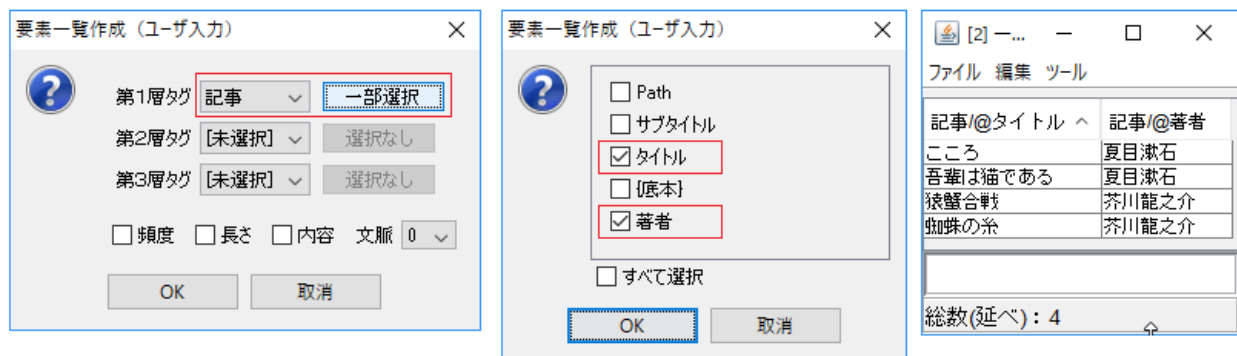
6.1 基本的な利用方法 [↑]

[ツール]⇒[一覧]⇒[ユーザ入力]で、言語資料に付与されているアノテーション内容の集計を行います。

アノテーションは、タグによって記述されているため、タグを指定して集計することになります。例えば、『青空文庫』サンプルでは、一つの作品に対して、「記事」というタグが付与されています。タグはいくつかの属性を持つことができ、「記事」には作品名や著者名の属性があります。

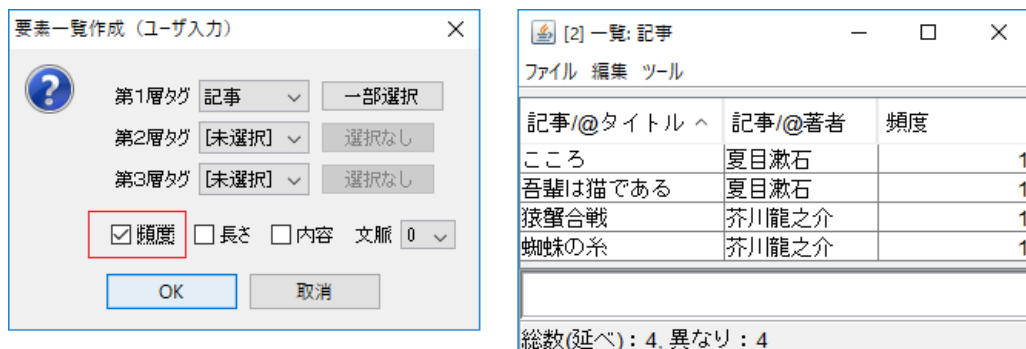
『青空文庫』サンプルの「記事」タグを使って、作品一覧を作成する手順は、次のとおりです。

1. [ツール]⇒[一覧]⇒[ユーザ入力]で設定用のウィンドウを起動して、下図（左）のように「第1層タグ」のところに、「記事」を設定して下さい。
2. 選択メニューの右のボタンを押すと、下図（中央）のウィンドウが現れるので、「タイトル」「著者」にチェックを入れて下さい。
3. 二つのウィンドウの「OK」ボタンを押すと、下図（右）の記事一覧表が作成されます。



6.2 「頻度」オプション [↑]

「頻度」オプションをチェックすると、一覧の各項目の出現頻度を計測することができます。下の図は、「記事」の頻度を表示したものです。『青空文庫』サンプルには、作品は重複して登録されていないため、当然、各作品の頻度は1になります。



同様に、rタグ（ルビ）に対して、実行したのが次の図です。左図はrタグの属性rtを選択して表示したものです。rt属性には、ルビ本体が記述されているので、頻度付きのルビの一覧を作成することができます。

要素一覧作成 (ユーザ入力)

第1層タグ 全選択

第2層タグ [未選択] 選択なし

第3層タグ [未選択] 選択なし

☒ 頻度 ☐ 長さ ☐ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

☒ r

☒ すべて選択

OK 取消

r/{r}	頻度
い	183
うち	178
か	134
ま	110
ぎ	109
よ	104
134	
総数(延べ): 13932, 異なり: 4746	

一方、右図は属性を選択しないで表示したものです。この場合、rタグの総数を計測することになります。

要素一覧作成 (ユーザ入力)

第1層タグ 選択なし

第2層タグ [未選択] 選択なし

第3層タグ [未選択] 選択なし

☒ 頻度 ☐ 長さ ☐ 内容 文脈 0

OK 取消

[1] 一覧: r

ファイル 編集 ツール

頻度

13932

総数(延べ): 13932, 異なり: 1

6.3 「第x層タグ」の設定

タグは、「第1層タグ」「第2層タグ」「第3層タグ」に複数指定することにより、タグ間の包含関係を考慮した一覧の作成が可能です。下の図は、「第1層タグ」に「記事」タグ、「第2層タグ」にrタグを指定することにより、「記事」に含まれるルビの数を計測しています。

要素一覧作成 (ユーザ入力)

第1層タグ 一部選択

第2層タグ 選択なし

第3層タグ [未選択] 選択なし

☒ 頻度 ☐ 長さ ☐ 内容 文脈 0

OK 取消

[3] 一覧: 記事/r

ファイル 編集 ツール

記事/@P...	記事/@タイ...	記事/@著者	頻度
/aozora_sam...	こころ	夏目漱石	4567
/aozora_sam...	吾輩は猫であ...	夏目漱石	9214
/aozora_sam...	猿蟹合戦	芥川龍之介	83
/aozora_sam...	蜘蛛の糸	芥川龍之介	68

総数(延べ): 13932, 異なり: 4

「頻度」は最下層のタグを対象に計測します。上の例の場合は、rタグの頻度を「記事」ごとに計測することになります。

6.4 「長さ」オプション

「長さ」オプションは、タグでマークアップされている文字列の長さを計測します。この際、マークアップされている文字列の中に含まれるタグや空白文字は、すべて長さ0として計測されます。

次の例は、「記事」タグでマークアップされている文字列（『青空文庫』サンプルの場合は一つの作品）に含まれる文字数を計測することになります。

要素一覧作成 (ユーザ入力)

第1層タグ 一部選択

第2層タグ [未選択] 選択なし

第3層タグ [未選択] 選択なし

☐ 頻度 ☒ 長さ ☐ 内容 文脈 0

OK 取消

[5] 一覧: 記事

ファイル 編集 ツール

記事/@Path	記事/@タイトル	記事/@著者	記事%文字数
/aozora_samp...	こころ	夏目漱石	161509
/aozora_samp...	吾輩は猫である	夏目漱石	319370
/aozora_samp...	猿蟹合戦	芥川龍之介	2739
/aozora_samp...	蜘蛛の糸	芥川龍之介	3389

総数(延べ): 4

6.5 「内容」オプション ⁺

「内容」オプションは、タグでマークアップされている文字列のための列を集計結果に追加します。

次の例は、rタグでマークアップされている文字列、つまり、ルビをつけられている文字列とルビをペアで集計しています。

要素一覧作成 (ユーザ入力)

第1層タグ: r 全選択

第2層タグ: [未選択] 選択なし

第3層タグ: [未選択] 選択なし

☒ 頻度 ☐ 長さ ☒ 内容 文脈 0

OK 取消

[6] 一覧: r

ファイル 編集 ツール

r/[@{rt}]	r%内容	頻度
はい	這入	77
ま	間	77
あと	後	72
だいぶ	大分	72
わたくし	私	71
おおい	大	70
い	好	69

77

総数(延べ): 13932, 異なり: 5975

6.6 「文脈」オプション ⁺

「文脈」オプションは、指定したタグのうち、最下層のタグに関して、後続するnタグ分の情報を集計結果に追加します。なお、nは「文脈」オプションで指定した値です。

例えば、『青空文庫』サンプル（形態素解析結果付き）のmorphタグを使って、単語bigramを作成してみます。morphタグは「単語」をマークアップするためのタグです。「文脈」オプションの値は1とします。また、あわせて、「頻度」オプションもチェックします。この場合、後続する1単語をペアにして計測することになるので、bigramが得られることになります。ただし、『ひまわり』の内部では、作品の最後の単語の次の単語は、次の作品の先頭の単語として、記述されているため、一部不要なデータbigramの定義に沿わないデータが含まれることに注意して下さい。

要素一覧作成 (ユーザ入力)

第1層タグ: morph 一部選択

第2層タグ: [未選択] 選択なし

第3層タグ: [未選択] 選択なし

☒ 頻度 ☐ 長さ ☐ 内容 文脈 1

OK 取消

要素一覧作成 (ユーザ入力)

☐ {TEXT}

☒ 品詞
☒ 品詞細分類1
☒ 品詞細分類2
☒ 品詞細分類3
☒ 基本形
☒ 活用型

☐ 活用形
☐ 発音
☐ 読み

☐ すべて選択

OK 取消

結果は、次のとおりです（一部の列のみ表示）。これを見ると、最も出現頻度の多いのは、「た」 + 「。」であることがわかります。morph[0], morph[1]がそれぞれ1番目、2番目の単語を表します。

[7] 一覧: morph

ファイル編集ツール

morph/@品詞	morph/@基本形	morph[1]/@品詞	morph[1]/@基本形	頻度 ▾
助動詞	た	記号	。	3829 ^
助詞	て	動詞	いる	2479
記号	」	記号	「	2151
動詞	する	助詞	て	1747
助詞	で	記号	、	1488
名詞	私	助詞	は	1385 v

総数(延べ): 315568, 異なり: 92901

6.7 外部アノテーション結果の表示 ±

形態素解析結果など、外部アノテーションが施されている資料（言語資料の選択時に「外部DBあり」の資料）では、SHIFTキーを押しながら、検索結果をダブルクリックすると、当該の作品の外部アノテーション結果が一覧表示されます。

次の例は、『青空文庫』サンプル（形態素解析付き）で、「我輩」を検索し、その中の一つをSHIFT+ダブルクリックした結果です。1行1形態素で、「_TEXT」列が本文に相当します。

[11] 一覧

ファイル編集ツール

SER.NO. ^	_TEXT	品詞	品詞細分類 1	品詞細分類 2	品詞細分類 3	活用型	活用形
00000001	一	名詞	数				
00000002		記号	空白				
00000003	吾輩	名詞	代名詞	一般			
00000004	は	助詞	係助詞				
00000005	猫	名詞	一般				
00000006	で	助動詞				特殊・ダ	連用形
00000007	ある	助動詞				五段・ラ行ア...	基本形
00000008	。	記号	句点				
00000009	名前	名詞	一般				
00000010	は	助詞	係助詞				
00000011	まだ	副詞	助詞類接続				
00000012	無い	形容詞	自立			形容詞・アウ...	基本形
00000013	。	記号	句点				
00000014		記号	空白				
00000015	どこ	名詞	代名詞	一般			
00000016	で	助詞	格助詞	一般			
00000017	生れ	動詞	自立			一段	連用形

00000003

総数(延べ): 206322