

全文検索システム『ひまわり』/利用者マニュアル/1_6/5. 検索結果を集計する

[Top](#) / [全文検索システム『ひまわり』](#) / [利用者マニュアル](#) / [1_6](#) / 5. 検索結果を集計する

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](#)

5. 検索結果を集計する [↑]

5.1 頻度を計測する [↑]

次の図は、『青空文庫』サンプルから「これ」を検索した結果です。ここでは、作品別の出現頻度を求めてみます。

頻度の計測は、列を指定して行います。作品別の頻度を計測する場合、「タイトル」列のいずれかのセルを選択し、右クリック⇒「統計」を実行します。

全文検索システムひまわり - [『青空文庫』サンプル(形態素解析結果付き)] - config_aozora_sample.sd.xml

ファイル 編集 ツール ヘルプ

検索文字列 フィルタ コーパス 検索オプション

本文 ▼ これ

前文脈

後文脈

で終る ▼

で始まる ▼

検索

字体変換

クリア

no	前文脈	キー	後文脈	Path	タイトル	著者	品詞
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...	こころ	吾輩は猫...	詞
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
3	弾くところです」	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
4	い話があるかい」	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
6	見当がつかない」	これ	からいよいよ弾くところ	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
7	めちゃんとお困ります。	これ	からがいよいよ佳境に	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
8	うと云うんです。さあ	これ	からがいよいよ失恋に	/aozora_s...	吾輩は猫...	吾輩は猫...	詞
9	はすこぶる不慥だよ。	これ	からがいよいよ巧妙な	/aozora_s...	吾輩は猫...	夏目漱石	副詞
10	充分あらわれている。	これ	からが化物の記述だ。	/aozora_s...	吾輩は猫...	夏目漱石	副詞
11	か両君能く聞き給え、	これ	からが結論だぜ。一	/aozora_s...	吾輩は猫...	夏目漱石	副詞
12	と一と息ついた。「	これ	からが聞きどころです	/aozora_s...	吾輩は猫...	夏目漱石	副詞
13	んだ。「まだです。	これ	からが面白いところで	/aozora_s...	吾輩は猫...	夏目漱石	副詞

1

こころ

検索総数: 597

計測結果は、次のようになります。

タイトル	頻度
吾輩は猫である	476
こころ	116
蜘蛛の糸	4
猿蟹合戦	1

総数(延べ): 597, 異なり: 4

複数のセルを選択すると、選択した列の値を組にして頻度を計測することができます。次の例は、タイトルと作品を組にした場合の結果です。

タイトル	著者	頻度
吾輩は猫...	夏目漱石	476
こころ	夏目漱石	116
蜘蛛の糸	芥川龍之介	4
猿蟹合戦	芥川龍之介	1

総数(延べ): 597, 異なり: 4

5.2 検索結果・集計結果を編集する

正規表現置換により、検索結果、集計結果を編集します。

ここでは、年月日表示から年表示にする例を示します。使用した資料は、国会会議録パッケージです。例えば、この処理により、年ごとの集計が容易になります。

まず、「開催年月日」列のセルを右クリックし、「置換」を実行します。

全文検索システムひまわり - [国会会議録_20140327_rev20170612] - config_kokkai_honkaigi.xml

ファイル 編集 ツール ヘルプ

検索文字列 フィルタ コーパス 検索オプション

討議部分 ▼ あの

前文脈

後文脈

で終る ▼

で始まる ▼

検索

字体変換

クリア

号	発言者	発言者(...)	肩書き	生年	開催日	文字数(...)	文字数(...)	URL
36	星野芳樹	星野芳樹		1909	1949-05-27	80084	80834	http://kokk...
02	野村哲郎	野村哲郎		1943	2012-11	36679	36944	http://kokk...
05	倉石忠雄	倉石忠雄	国務大臣	1900	1956-01	41099	41353	http://kokk...
07	受田新吉	受田新吉		1910	1957-02	10280	11328	http://kokk...
33	徳田球一	徳田球一		1894	1948-03	36391	36591	http://kokk...
02	吉田茂	吉田茂	国務大臣	1878	1951-08	29803	29941	http://kokk...
34	大山郁夫	大山郁夫		1880	1952-04	45692	48674	http://kokk...
12	久保等	久保等		1916	1965-12	88146	88788	http://kokk...
02	野坂参三	野坂参三		1892	1977-01-31	30891	35141	http://kokk...
12	野坂参三	野坂参三		1892	1950-01-25	39422	39567	http://kokk...
25	細迫兼光	細迫兼光		1896	1956-03-22	49671	50621	http://kokk...
01	神山茂夫	神山茂夫		1905	1949-12-04	43410	43979	http://kokk...
10	春日一幸	春日一幸		1910	1974-01-24	63223	64039	http://kokk...

1

1949-05-27

検索総数:7436

置換の設定は、置換元（正規表現，「-.*」），置換先を指定します。この場合，「-」以降の文字列を削除することにより，年表示にしています。なお，置換の処理は，Javaの[String#replaceAll](#)で行っています。後方参照についても利用可能です。

置換（正規表現）

検索する文字列 -.*

置換後の文字列

OK Cancel

結果は次のとおりです。新しいウィンドウが生成されて，置換結果が表示されます。

[1] 頻度：no,前文脈,キ-,後文脈,議院,回,会議名,号,発言者,発言者(正規化),肩書き,生年,開催日,文字数(討議),文字...

ファイル 編集 ツール

号	発言者	発言者(...)	肩書き	生年	開催日	文字数(...)	文字数(...)	URL
36	星野芳樹	星野芳樹		1909	1949	80084	80834	http://kokk...
02	野村哲郎	野村哲郎		1943	2012	36679	36944	http://kokk...
05	倉石忠雄	倉石忠雄	国務大臣	1900	1956	41099	41353	http://kokk...
07	受田新吉	受田新吉		1910	1957	10280	11328	http://kokk...
33	徳田球一	徳田球一		1894	1948	36391	36591	http://kokk...
02	吉田茂	吉田茂	国務大臣	1878	1951	29803	29941	http://kokk...
34	大山郁夫	大山郁夫		1880	1952	45692	48674	http://kokk...
12	久保等	久保等		1916	1965	88146	88788	http://kokk...
02	野坂参三	野坂参三		1892	1977	30891	35141	http://kokk...
12	野坂参三	野坂参三		1892	1950	39422	39567	http://kokk...
25	細迫兼光	細迫兼光		1896	1956	49671	50621	http://kokk...
01	神山茂夫	神山茂夫		1905	1949	43410	43979	http://kokk...

1949

総数(延べ)：7436

5.3 集計結果を合算する [†]

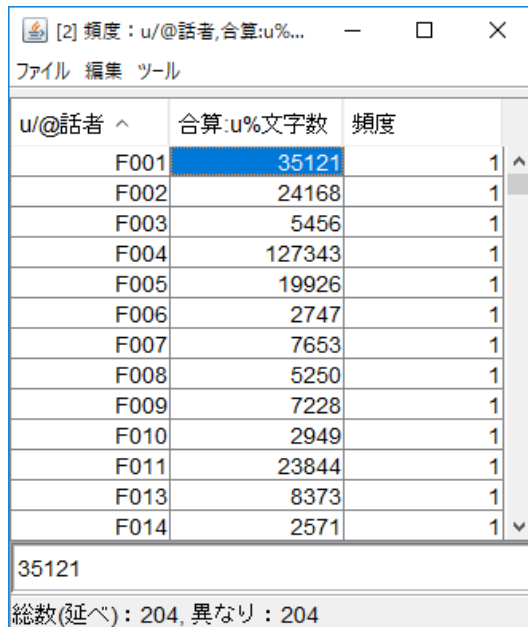
セルの値が数値の場合、それらを合算する機能です。名大会話コーパスパッケージを使って、話者ごとの発話文字数を計測してみます。

まず、各発話の文字数を「アノテーション内容の集計」機能（[ツール]⇒[一覧]⇒[ユーザ入力]）で求めます。一つの発話は、uタグでマークアップされています。さらに、発話者の名前を表示するため、「話者」の属性をチェックします。また、発話の文字数とその頻度も表示するように、「頻度」「長さ」もチェックします。頻度を表示するのは、同じ文字数の発話が複数存在する可能性があるからです。

集計結果は、結果は、次のとおりです。例えば、先頭行は、話者「F001」の発話のうち、文字数が13だったものが、107回あったことを表します。

[3] 一覧: u		
ファイル 編集 ツール		
u/@話者 ^	u%文字数	頻度
F001	13	107
F001	112	1
F001	121	2
F001	160	1
F001	102	1
F002	9	43
F002	7	37
F002	8	41
F002	5	37
F002	6	23
F002	3	122
F002	4	50
F002	1	2
F001		
総数(延べ): 173296, 異なり: 12663		

最後に、合算したい列のセル（「合算:u%文字数」）を選び、[編集]⇒[合算]を実行します。



u/@話者	合算: u%文字数	頻度
F001	35121	1
F002	24168	1
F003	5456	1
F004	127343	1
F005	19926	1
F006	2747	1
F007	7653	1
F008	5250	1
F009	7228	1
F010	2949	1
F011	23844	1
F013	8373	1
F014	2571	1

35121

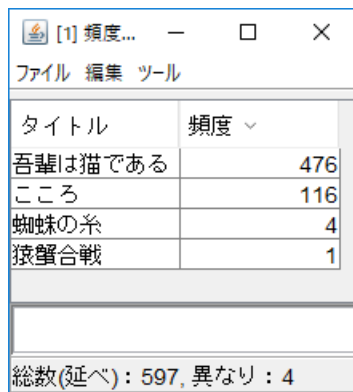
総数(延べ): 204, 異なり: 204

合算では、合算する列と「頻度」列を除くすべての列の値が同じ行の値が合算されます。上の例の場合は、「話者」列の値が同じ場合、「合算: u%文字数」列の値を合算します。合算する際は、「頻度」列の値を考慮し、先ほど例示した先頭行の場合、13×107文字として、計算されます。

5.4 集計結果を結合する [↑]

この機能は、二つの集計結果を結合する機能です。

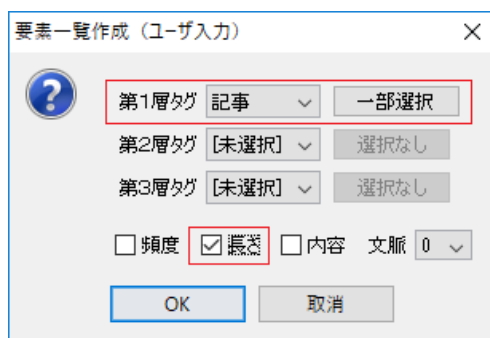
5.1では、『青空文庫』サンプルを例として、「これ」の出現頻度を作品別に集計しましたが、作品ごとに総文字数が異なっているため、直接比較することはできません。ここでは、「結合」機能を使って、出現頻度の集計結果（表a）に、作品ごとの総文字数を追加してみます。



タイトル	頻度
吾輩は猫である	476
こころ	116
蜘蛛の糸	4
猿蟹合戦	1

総数(延べ): 597, 異なり: 4

作品ごとの総文字数を求めるには、次のように、「アノテーション内容の集計」機能（[ツール]⇒[一覧]⇒[ユーザ入力]）で求めます。



要素一覧作成 (ユーザ入力)

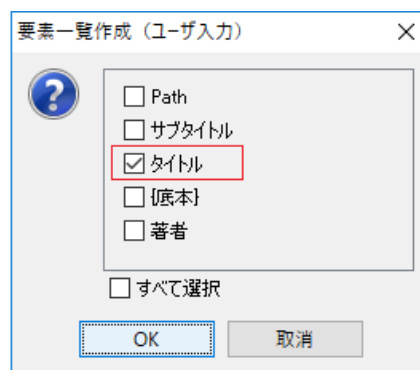
第1層タグ: 記事 (一部選択)

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 ☒ 異なり ☐ 内容 文脈 0

OK 取消



要素一覧作成 (ユーザ入力)

☐ Path

☐ サブタイトル

☒ タイトル

☐ 〔原本〕

☐ 著者

☐ すべて選択

OK 取消

結果は、次のとおりです（表b）。

[5] 一覧: ー □ ×	
ファイル 編集 ツール	
記事/@タイトル ^	記事%文字数
こころ	161509
吾輩は猫である	319370
猿蟹合戦	2739
蜘蛛の糸	3389
総数(延べ) : 4	

表aに表bの頻度列（総単語数）を結合します。結合には、まず、表bから結合したい列とキーとなる列を指定します。

キーとなる列とは、二つの表の値を結合する時、基準となる列です。ここでは、作品名がキーとなるので、「記事/@タイトル」列がキーとなります。結合したい列は「記事%文字数」列なので、「記事/@タイトル」「記事%文字数」列のセル（どれでもよい）を選択して、[編集]⇒[コピー（列名を含む）]を実行します。

[5] 一覧: ー □ ×	
ファイル 編集 ツール	
記事/@	検索 Ctrl+F
こころ	置換 Ctrl+R
吾輩は	コピー Ctrl+C
猿蟹合	コピー(列名含む) Ctrl+Shift+C
蜘蛛の	結合 Ctrl+J
こころ	合算
総数(延べ) : 4	

次に、結合先の表aからキーとなる列のセルを選択します。この場合は「タイトル」列なので、表aで「タイトル」列のセルを一つ選択した後、[編集]⇒[結合]を実行して下さい。結果は、次のようになります。

[9] 頻度: タイ... ー □ ×		
ファイル 編集 ツール		
タイトル	記事%文字数	頻度 ▾
吾輩は猫...	319370	705
こころ	161509	258
蜘蛛の糸	3389	11
猿蟹合戦	2739	5
総数(延べ) : 979, 異なり : 4		