

# digitdemog

## Perl製UNIXシェル型コマンド

2022-03-03 fri 下野寿之

# 初めに

- `digitdemog`という便利なコマンドを説明する。
- 本文書は文量が多い。初めて読むと、意味が分かりにくくいであろう。
- しかし、このコマンドを使うと、便利さに気づくであろう。
- 使いながら必要な機能をできるだけ、汎用性と再利用性を高めて実装したからである。
- 端末で、出力の行数が非常に多いことも頻繁に起こりうる。しかし、仕組みを理解すれば、短く把握が容易な形で出力させることは簡単であろう。

# 前提としたスキル

- Unixコマンドが使えること
  - オプション指定などで、いろいろな操作をすること。
- bashやzshを使っていること
  - パイプ( | )などが使えること
- 文字コードを知っていること
  - アスキーコードとUnicode。
- 正規表現を知っていること
  - 文字クラス
  - Perl特有のものも本文書に含まれるかも知れない。
- Perlのモジュール(ライブラリ)をインストール可能な事。

# 概要

- `digitdemog` は端末上で、コマンドとして使う。
  - 名前の由来は 桁(digit)の人口動態(demography)である。
  - Perlのモジュール `App::digitdemog` がこのコマンドを提供する。
- インストール方法:
  - ① Perlのモジュールの通常のインストール方法を用いる。
  - ② すなわち、`cpan App::digitdemog` を実行する。
  - ③ インストール前に戻せるように、`cpan`でなく `cpanm` が推奨される。
- `digitdemog` のマニュアル参照の方法はいくつかある。
  - ① `digitdemog --help` を実行して参照。
  - ② `man App::digitdemog`
  - ③ `perldoc App::digitdemog`

# 機能

- テキストデータの全部の行を読んで、出現した異なる文字の全てについて、各行の左から何桁目(何文字目)に、その文字が何度出現したかを、二元分割表(クロス集計表)として出力する。
- その分割表は、縦方向にそれらの異なる各文字に対応し、横方向に桁位置が対応する。
- その分割表の右側には次を配置する。
  1. 出現した各文字  $c$
  2. その文字  $c$  の文字コード (ユニコードにおける文字コード)
  3.  $c$  の出現度数(頻度)の全桁における和 (文書全体における  $c$  の頻度)
  4. その文字を含む各行についての行全体の文字列の内の、いくつか

# オプション機能

下記のそれぞれを選んで有効化できる。

1. 1行目だけを読み飛ばす。
2. 全く同じ行が現れたら、読み飛ばす。
3. 文字を、数字や平仮名などでグルーピング。
4. 元の各行全体の例を、約何個取り出すか指定。
5. 出力表の中の頻度に数値 $r$ が表示されれば、 $u$ 行の入力の内 $r$ 個の行の組み合わせ( ${}_u C_r$ 個考えられる内の1個)が対応している。同じ $r$ の表示のうち全く同じ組み合わせに対応するものが多数有る場合は、その数 $r$ にピリオド(.)を付与して表示する。

# 有用性

- 主に表形式データの各列の検査に有用。
  - 元の機能は、検査対象は各行である。
  - 従って、気になる列に対し1列ずつ検査を進める。
- 含まれる文字の様子が、全体的に把握できる。
  - 数字だけか他の記号を使ってないか。
  - アスキー文字だけか、×や顔文字を使っていないか。
  - 平仮名・カタカナ・漢字・記号のどれを使っているか。
- 十数桁以内の符号で、規則性があれば検査可能。
  - 電話番号でハイフンや括弧を使っているか。
  - 図書のISBN(数13桁)に現れるハイフンの様子。
- 日付などの書式が解読できる。
- 例外的な不規則な箇所があれば、検出が容易。
  - 規則性を要求されるデータの形式に新しい提案が可能。

# 利用例1

*digitdemog* のコマンドの実行結果から  
いろいろな知見が得られる。  
その知見の読み取れる出力表は  
コピペなどにより、エクセルなどに  
保存も出来て使い勝手も良い。

「TRC新刊図書オープンデータ」(毎週更新)

[https://www.trc.co.jp/trc\\_opendata/](https://www.trc.co.jp/trc_opendata/)

# 参考: 例に用いたデータについて

- 「TRC新刊図書オープンデータ」を用いた。  
[https://www.trc.co.jp/trc\\_opendata/](https://www.trc.co.jp/trc_opendata/)
- 毎週更新される18列のデータである。
- 各列の性質は、以下の通り：

```
> colsummary -m0 -j -g2 TRCOpenBibData_20230225.txt
列番 異なる値 値の範囲 最頻値 頻度(重複) 衍数
1 1560 1978-4-00-027249-0~978-4-9911149-4-6 1978-4-86240-205-9 4|3(2)|1(1557) 0|17
2 1559 #シンFIRE論~黒鳶の聖者 5 やさしくたって満足したい!クラシック名曲アルバム|世界でいちばんやさしい教養の教科書 3|2(6)|1(1552) 1~53
3 701 |0・1・2歳児 文例たっぷり!!~魔界の常識で生きてたら、気付けば人類最強になっていた |文部科学省後援 811|9~2(16)|1(673) 0|2~121
4 1296 |Bryan F.J.Manly 編~齋藤寛 著 |協同教育研究会 編 148|13~2(48)|1(1227) 0|3~40
5 383 |.suke イラスト~黒冴 イラスト |大岩秀樹 著 1155|6~2(18)|1(360) 0|4~25
6 51 |13訂版~縮刷版 |改訂版 1440|21~2(8)|1(32) 0|2~14
7 455 |ALBA~鹿砦社 KADOKAWA|講談社 180|58~2(64)|1(243) 0|2~31
8 54 |Gakken~集英社 |星雲社 1389|29~2(6)|1(31) 0|2~15
9 8 2022.11~2023.4|c2023 2023.2|2023.3 975|531~17|1(3) 4~7
10 470 |10, 15, 1976p~p97~192 |1冊(ページ付なし) 124|48~2(80)|1(216) 0|2~13
11 38 |15cm~31cm 19cm|21cm 395|253~2(2)|1(16) 0|4~7
12 6 |CD-ROM(1枚 12cm)~録音ディスク(2枚 12cm) |録音ディスク(1枚 12cm) 1551|6~2|1 0|15~19
13 526 |&FLOWERフラワーコミックスα~青森県の教員採用試験過去問シリーズ 12 |角川コミックス・エース 778|18~2(33)|1(441) 0|3~33
14 56 |&.Emo comics~関東 03 |レベル別問題集シリーズ 1480|6(2)~2(8)|1(38) 0|2~25
15 3 |ALBA GREEN BOOK~辻番奮闘記 5 |ALBA GREEN BOOK 1565|1(2) 0|7~15
16 151 |17階級徹底調査でまるわかり。リングをおもしろくするヒーローたち~鹿児島県統計書 |公害総論 1415|2(2)|1(148) 0|1~82
17 193 ¥1000~¥9900 ¥1600|¥1800 76|73~2(23)|1(79) 4~6
18 2 |3巻セット¥9000 |3巻セット¥9000 1561|6 0|10
1,567 line(s) read; 0.081857 seconds (colsummary)
>
```

# 書誌情報の本の「ISBN」のデータから読み取れること-1

<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	<b>16</b>	<b>17</b>	<b>18</b>	<b>char</b>	<b>code</b>	<b>freq</b>	<b>example1..example</b>		
0	0	0	1563	0	1563	0	0	477	493	358	181	52	2	0	1563	0	0	0	'-	U+2D	6252	978-4-00-027249-0\n1978-4-9911149-4-6\n		
0	0	0	0	0	0	392	140	88	93	242	215	144	196	149	0	158	0	0	'0'	U+30	1817	978-4-00-027249-0\n1978-4-09-872013-2\n		
0	0	0	0	0	0	85	135	86	135	206	226	134	152	168	0	143	0	0	'1'	U+31	1470	978-4-10-603071-0\n1978-4-19-901093-4\n		
0	0	0	0	0	0	106	109	80	107	140	204	175	185	178	0	153	0	0	'2'	U+32	1437	978-4-251-07862-9\n1978-4-299-04076-3\n		
0	0	0	0	0	0	121	117	119	95	196	191	170	134	158	0	147	0	0	'3'	U+33	1448	978-4-300-10114-8\n1978-4-398-29674-0\n		
0	0	0	0	0	0	1563	0	97	263	151	75	65	142	199	141	168	0	162	0	0	'4'	U+34	3026	978-4-00-027249-0\n1978-4-9911149-4-6\n
0	0	0	0	0	0	151	95	136	130	52	119	147	166	148	0	135	0	0	'5'	U+35	1279	978-4-502-45271-0\n1978-4-596-76929-9\n		
0	0	0	0	0	0	18	273	112	144	42	77	159	157	168	0	181	0	0	'6'	U+36	1331	978-4-620-79462-4\n1978-4-651-20312-6\n		
0	1563	0	0	0	0	181	91	111	88	82	96	118	136	141	0	148	0	0	'7'	U+37	2755	978-4-00-027249-0\n1978-4-9911149-4-6\n		
0	0	1563	0	0	0	358	113	80	97	105	53	133	137	151	0	162	0	0	'8'	U+38	2952	978-4-00-027249-0\n1978-4-9911149-4-6\n		
1563	0	0	0	0	0	54	227	123	106	75	59	132	157	134	0	174	0	0	'9'	U+39	2804	978-4-00-027249-0\n1978-4-9911149-4-6\n		
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1563	0	"\n"	U+0A	1567	\n			
0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	\$	end	1567	\n			

- 全ての1563行が、17文字+(改行文字)である。

17桁のISBNコードのパターンがある程度読み取れる(ただし特定の1週間である)。

- 全ての行が「978-4-」で始まる。

- ハイフンの位置は決まっているとは限らない。

- 0始まりの、3桁目、5桁目、15桁目は、全行においてハイフンである。
- しかし、それ以外にも9桁目は、493個の行(約31.5%)だけがハイフンである。
- 8桁目、10桁目、11桁目、12桁目、13桁目も、同様に、全てでは無いが一部の行においてそうなる。

- 0始まりの6桁目から14桁目にかけて、数字の0~9が現れる。

- ハイフンが現れることもある。
- 数値の分布は、各桁において、大体一様なことであれば、頻度がばらつくこともある。

- オプションの-.0により、頻度に付くピリオド(.)を抑制。
- この表では1563の数値にしか付かなかったため、横幅を縮めることを優先した。

## 書誌情報の本の「大きさ」のデータから読み取れること:

> cut -f11 TRCOpenBibData_20230225.txt   digitdemog -0- -n0											
<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>char</u>	<u>code</u>	<u>freq</u>	<u>example.. example</u>
-	143	-	-	6	-	-	-	'0'	U+30	149	20cm 20x20cm
656	304	-	1	5	-	-	-	'1'	U+31	966	18x26cm 18cm
663	74	-	33	2	-	-	-	'2'	U+32	772	26cm 21cm
124	16	-	7	2	-	-	-	'3'	U+33	149	31cm
-	15	-	-	2	-	-	-	'4'	U+34	17	24cm 24x26cm
-	139	-	-	1	-	-	-	'5'	U+35	140	15cm
-	204	-	-	20	-	-	-	'6'	U+36	224	26cm
-	29	-	-	2	-	-	-	'7'	U+37	31	27cm 17x22cm
-	91	-	-	-	-	-	-	'8'	U+38	91	18x26cm 18cm
-	428	-	-	1	-	-	-	'9'	U+39	429	19cm
-	-	1402.	-	-	41.	-	-	'c'	U+63	1443	26cm 18cm
-	-	-	1402.	-	-	41.	-	'm'	U+6D	1443	26cm 18cm
-	-	41.	-	-	-	-	-	'x'	U+D7	41	18x26cm 17x22cm
124	-	-	-	1402.	-	-	41.	\$	end	1567	

- 全部で1567行(最終行の黄色の\$は文末を意味する。)
- 全ての行は、¥n(U+0A)で終わる。
- 空文字列の行が124個。それ以外は全てcmの2文字で終わる。
- 数字やcmの他にxではない×(U+D7)が使われている。(エックスxはU+78)
  - 見た目が似た文字の、文字コードもU+の表示で区別できるので、便利。
- 右側に|で区切って、行全体の具体例をいくつか表示。
  - digitdemogの動作に疑義が生じても、具体例で確認できる。
- オプションの-0-で頻度0はハイフンで表示。
- オプションの-n0で改行文字はカウントの対象外としている。

## 書誌情報の本の「ISBN」のデータから読み取れること-2

	<code>&gt; digitdemog &lt;(cut -f1 TRCOpenBibData_20230225.txt ) -e'\d{8}' -e'\d{7}' -e'\d{6}' -e'\d{5}' -e'\d{4}' -e'\d{3}' -e'\d{2}'</code>	<code>char</code>	<code>code</code>	<code>freq</code>	<code>example..example</code>
0	1 2 3 4 5 6 7 8 9 10	\d{8}	---	0	
0	0 0 0 0 0 0 0 0 0 0	\d{7}	---	2	978-4-9911149-3-9\n 978-4-9911149-4-6\n
0	0 0 0 2 0 0 0 0 0 0	\d{6}	---	529	978-4-901794-26-8\n 978-4-944185-19-1\n
0	0 0 0 52. 0 477. 0 0 0 0	\d{5}	---	674	978-4-86053-144-7\n 978-4-89831-969-7\n
0	0 0 0 181. 0 493. 0 0 0 0	\d{4}	---	716	978-4-7503-5527-6\n 978-4-8471-4938-2\n
1563.	0 0 0 493. 0 181. 0 0 0 0	\d{3}	---	2237	978-4-00-027249-0\n 978-4-9911149-4-6\n
0	0 0 0 477. 0 52. 0 0 0 0	\d{2}	---	529	978-4-00-027249-0\n 978-4-19-901093-4\n
0	1563. 0 1563. 0 1563. 0 1563. 0 0 0 0	'-	U+2D	6252	978-4-00-027249-0\n 978-4-9911149-4-6\n
0	0 0 0 0 0 0 0 0 0 0	'0'	U+30	158	978-4-00-027249-0\n 978-4-910815-06-0\n
0	0 0 0 0 0 0 0 0 0 0	'1'	U+31	143	978-4-01-093297-1\n 978-4-944185-19-1\n
0	0 0 0 0 0 0 0 0 0 0	'2'	U+32	153	978-4-00-602349-2\n 978-4-910539-07-2\n
0	0 0 0 0 0 1 0 0 0 0	'3'	U+33	148	978-4-9911149-3-9\n
0	0 1563. 0 0 0 1 0 0 0 0	'4'	U+34	1726	978-4-00-027249-0\n 978-4-9911149-4-6\n
0	0 0 0 0 0 0 0 0 0 0	'5'	U+35	135	978-4-00-312281-5\n 978-4-910803-13-5\n
0	0 0 0 0 0 0 0 0 0 0	'6'	U+36	181	978-4-01-035054-6\n 978-4-9911149-4-6\n
0	0 0 0 0 0 0 0 0 0 0	'7'	U+37	148	978-4-01-093729-7\n 978-4-910603-11-7\n
0	0 0 0 0 0 0 0 0 0 0	'8'	U+38	162	978-4-00-372510-8\n 978-4-924523-38-8\n
0	0 0 0 0 0 0 0 0 0 0	'9'	U+39	174	978-4-01-035053-9\n 978-4-9911149-3-9\n
4.	0 0 0 0 0 0 0 0 0 0	"\n"	U+0A	1567	\n
0	4. 0 0 0 0 0 0 0 0 0	\$	end	1567	\n

- ISBNの列に対して「パターン」をいくつか手入力した。

- 数値のまとまりを、「8桁以上」「7桁以上」..「2桁以上」で見ている。

- この逆に指定すると、想定したとおりの動作にはならない。

- **数値と反復回数を表す正規表現**の表記の仕方について、知っていることを前提とする。

- 空文字列の4行以外は、数値(5回)とハイフン(4回)が交互に現れる。

- 各行で、数値の出現の3回目と4回目の出現について、桁数は合計が8桁。

- 出力表の.の付いた数で同じ頻度数であれば、それに対応する行は全ては、一致することを意味している。

- たとえば、52.と書かれた(0始まりのまとまり)4番目の「数6桁」と6番目の「数2桁」の両方に当てはまる行は全て同じ入力の行位置から来ていることを表している。181. や358. も同様。

- 入手したデータ上17桁だった「ISBN番号は13桁の数」について理解が深まる。

- 実際に確かめることは大事。何か誤解をしているという不安を払拭するため。

- データに規則性があり、データ上の不具合が無いこと(ただし空欄は4個あった)を確かめることが出来た。

- digitdemog単体でも有用だが、パターン指定により、的確なことが分かる。

- この出力だけでも記録すれば、後でいろいろなことを隨時確認できる。

- 上記の知見も**メモ取りは重要**(他人への説明に思い出すのに時間がかかるため)だが、**実務上は省略可能**。<sub>12</sub>

# 書誌情報の本の「ページ数等」から読み取れること：

cut -f10 TRCOpenBibData_20230225.txt   digitdemog -n0 -0- -.7   expandtab																	
0	1	2	3	4	5	6	7	8	9	10	11	12	13	char	code	freq	example1..example
-	-	4.	20	43	-	-	3.0	12.	1.00	-	-	-	-	'.'	U+20	83	9, 524p 5, 78p
-	-	48.	-	-	-	-	-	-	-	-	-	-	'('	U+28	48	1冊(ページ付なし)	
-	-	-	-	-	-	-	-	48.	-	-	-	-	')'	U+29	48	1冊(ページ付なし)	
-	4.	19	37	-	-	3.0	12.	1.00	-	-	-	-	-	','	U+2C	76	9, 524p 5, 78p
-	137	119	1	1.00000	2	4	3	-	-	-	-	-	-	'0'	U+30	267	100p 102p
532	136	167	1	4	18	9	7	2	13	-	-	-	-	'1'	U+31	889	175p 143p
445	167	93	4	8	8	6	1.00	1	-	11	-	-	-	'2'	U+32	744	228p 255p
233	116	139	3	6	8	4	-	1.0	-	1.00	-	-	-	'3'	U+33	511	336p 387p
84	111	101	-	2	5	2	2	1	1.000	-	-	-	-	'4'	U+34	309	431p 474p
36	183	127	3	2	8	8	1	-	-	-	-	-	-	'5'	U+35	368	513p 526, 82p
18	119	89	-	1	7	5	1.0	5	1.0000	1	1.	-	-	'6'	U+36	248	68p 63p
26	149	132	2	1.0	3	7	1	-	-	1.	-	-	-	'7'	U+37	322	79p 77p
21	128	101	-	1	5	4	3	1.000	-	-	-	-	-	'8'	U+38	264	80p 82p
38	125	145	1.00000	-	3	5	-	-	1.	-	-	-	-	'9'	U+39	318	95p
10	1	139	1163	3.00	3.	11	26	4.0	7	1.0000	13.	1.	-	'p'	U+70	1382	p145~280 p1327~1631
-	-	-	2	7	1.0	-	1.000	-	-	-	-	-	-	'~'	U+301C	11	p93~232 p97~192
-	-	-	-	-	-	-	-	48.	-	-	-	-	-	'し'	U+3057	48	1冊(ページ付なし)
-	-	-	-	-	-	-	-	48.	-	-	-	-	-	'な'	U+306A	48	1冊(ページ付なし)
-	-	-	-	-	48.	-	-	-	-	-	-	-	-	'ジ'	U+30B8	48	1冊(ページ付なし)
-	-	-	48.	-	-	-	-	-	-	-	-	-	-	'ペ'	U+30DA	48	1冊(ページ付なし)
-	-	-	-	48.	-	-	-	-	-	-	-	-	-	'一'	U+30FC	48	1冊(ページ付なし)
-	-	-	-	-	48.	-	-	-	-	-	-	-	-	'付'	U+4ED8	48	1冊(ページ付なし)
-	67	-	-	-	-	-	-	-	-	-	-	-	-	'冊'	U+518A	67	1冊 1冊(ページ付なし)
-	-	-	-	-	6.	-	-	-	-	-	-	-	-	'図'	U+56F3	6	292p 図版16p 541p 図版16p
-	-	-	-	1.000	-	-	-	-	-	-	-	-	-	'枚'	U+679A	1	23p 枚25~84
-	-	-	-	-	-	6.	-	-	-	-	-	-	-	'版'	U+7248	6	292p 図版16p 541p 図版16p
124	-	20	138	1157	3.00	3.	13	33	4.0	57	1.0000	13.	1.	\$	end	1567	

- 「数p」という行が多そうだが、いつもその形式とは限らない。
- 空文字列も10%近い124行。
- オプション-.7により、頻度1の所が1., 1.0, ... 1.00000 でどの同じ行から来たか識別可能にした。
- 1文字で収まらない決まったような言い回しがあるので、それを次ページで扱う。

## 書誌情報の本の「ページ数等」から読み取れること(2)：

```
> cut -f10 TRCOpenBibData_20230225.txt | digitdemog -n0 -0- -e'[0-9]+' -e'\(ページ付なし\)' -e'図版' -g2
0 1 2 3 4 5 6 7 8 char code freq example..example
1433 10. - 70 7 - 17 - - [0-9]+ --- 1537 175p|228p|195p|143p
- - 48. - - - - - \(ページ付なし\) --- 48 1冊(ページ付なし)
- - - 6. - - - - - 図版 --- 6 292p 図版16p|322p 図版16p|187p 図版16p|541p 図版16p
- - 67 - - 16. - - - - ' ' U+20 83 290, 7p|283, 96p|23p 枚25~84|19, 367p
- 60 - - 16. - - - - ',' U+2C 76 290, 7p|283, 96p|526, 82p|19, 367p
10. 1306 - - 44. 6. - 16. - 'p' U+70 1382 p145~280|p123~238|p103~188|p1327~1631
- - 10. - - 1. - - - - '～' U+301C 11 p145~280|p123~238|p103~188|p1327~1631
- 67 - - - - - - '冊' U+518A 67 1冊|1冊(ページ付なし)
- - - 1. - - - - - '枚' U+679A 1 23p 枚25~84
124 - 1318 48. 10. 44. 6. 1. 16. $ end 1567 1175p|195p
> cut -f10 TRCOpenBibData_20230225.txt | digitdemog -n0 -0- -e'[0-9]+' -e'\(ページ付なし\)' -e'図版' -g2.
0 1 2 3 4 5 6 7 8 char code freq example..example
1433 10. - 70 7 - 17 - - [0-9]+ --- 1537 10, 15, 1976p|27, 315, 12p|168, 8, 4p|23p 枚25~84
- - 48. - - - - - \(ページ付なし\) --- 48 1冊(ページ付なし)
- - - 6. - - - - - 図版 --- 6 292p 図版16p|322p 図版16p|187p 図版16p|541p 図版16p
- - 67 - - 16. - - - - ' ' U+20 83 10, 15, 1976p|27, 315, 12p|370, 122, 3p|168, 8, 4p
- 60 - - 16. - - - - ',' U+2C 76 10, 15, 1976p|27, 315, 12p|370, 122, 3p|168, 8, 4p
10. 1306 - - 44. 6. - 16. - 'p' U+70 1382 10, 15, 1976p|27, 315, 12p|370, 122, 3p|168, 8, 4p
- - 10. - - 1. - - - - '～' U+301C 11 23p 枚25~84|p145~280|p103~188
- 67 - - - - - - '冊' U+518A 67 1冊|1冊(ページ付なし)
- - - 1. - - - - - '枚' U+679A 1 23p 枚25~84
124 - 1318 48. 10. 44. 6. 1. 16. $ end 1567 10, 15, 1976p|27, 315, 12p|370, 122, 3p|168, 8, 4p
>
```

- オプションの-eでいくつも**正規表現**を指定して1文字と見なせるようにしてある。
  - その機能を使って、特定の言い回しを(前のページよりも)まとめて出力した。
- パターンがもっと明確に読み取れるようになった。
  - ページ数情報がどのように記載されたかが分かるので、もっと良い書式で記録することの提案が出来るであろう。
- この図の上半分と下半分は、オプションで-g2と-g2. と違う。
  - 上半分は**短めの例**、下半分は**長めの例**が現れやすい。
  - 出力各行において「左半分の各頻度」に対応している各入力行の文字列に対して、例については、前者は左(桁位置が0に近い方)、後者は右(桁位置が8に近い方)を優先した為。

# 書誌情報の本の「タイトル」から読み取れること：

cut -f2 TRCOpenBibData_20230225.txt   digitdemog -n0 -0- -e'[0-9]+ -e'[A-Za-z\ ]+ -e'[万-龍]+ -e'[あ-んア-ン]+ -e'[:ascii:]+ -.2  expandtab																								char	code	freq	example...example	
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	char	code	freq	example...example
46	60	110	111	123	100	63	74	52	26	26	17	5	10	5.	6.0	6	2.0	1.	2.	-	1	1.0	2	-	[0-9]+	---	849	10代のうちに考えておきたい「なぜ?」
94	124	92	117	78	64	56	49	23	18	15	10	7	5.	6.0	4	2.0	1.	2.	-	1	-	1	-	-	[A-Za-z\ ]+	---	769	TOEFLテストリスニング問題 IPALM TO P
900	285	648	242	400	191	228	116	116	68	59	32	37	15	18	9	6.	7	2	4.	3	3.	1	-	-	[万-龍]+	---	3390	腫瘍 内科 第31巻第2号(2023年2月)I 嘸
489	697	363	478	224	279	158	140	73	81	43	48	17	27	11	11	4	6.	4.	5	3.	3	-	-	-	[あ-んア-ン]+	---	3164	もうすぐやってくる尊皇攘夷思想のた&
4	44	11	36	25	20	11	10	6	6	2	4	2	2	4	-	1	-	1.0	-	-	-	-	-	-	[[:ascii:]]+	---	189	#シンFIRE論 I<ハーフ>物語
-	3	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'x'	U+D7	5	女子高生x日常に蠢く怪異!伸縮素材x口	
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'é'	U+E9	1	écriture新人作家・杉浦李奈の推論 8	
2	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	'“'	U+201C	3	“基盤モデル”で超進化するAI!“作りおき	
-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'”'	U+201D	3	“基盤モデル”で超進化するAI	
-	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'…'	U+2026	2	挙啓...殺し屋さんと結婚しました 5	
-	-	-	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'II'	U+2161	2	ロード・エルメロイII世の事件簿 10	
-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'V'	U+2164	1	STREET FIGHTER V CLIMAX ARTS+ZERO to	
-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'X'	U+2169	1	プロが教える!Final Cut Pro Xデジタル	
-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'★'	U+2605	1	K★STAR TXT4周年記念号	
-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'♥'	U+2661	2	ゆいたいむ has♡come!!	
-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'♥'	U+2665	2	制限♥解除	
-	11	15	7	11	6	6	4	2	2	2	1	1	-	-	1	-	-	-	-	-	-	-	-	'、'	U+3001	69	本屋、地元に生きる!ウツになつたら、	
-	1	3	2	3	3	4	4	3	2	2	1	1	1	-	-	1	1	-	-	-	-	-	-	'。'	U+3002	32	おとななちょ。	
-	1	-	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'々'	U+3005	4	威風堂々悪女 11	
-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'○'	U+3007	1	五〇歳からの勉強法	
2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'《'	U+300A	3	『一歩先を行く』リーダードリル<算数	
-	-	-	-	-	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'》'	U+300B	3	鉄人28号《オリジナル版》 11!『一歩	
18	5	6	5	2	2	2	3	2	2	4	1	1	-	-	-	-	-	-	-	-	-	-	-	'「'	U+300C	53	「しない」人になりなさい!「解釈のズ	
-	-	6	9	10	5	1	4	5	2	2	-	4	2	-	2	-	-	-	-	-	-	-	-	'」'	U+300D	53	「しない」人になりなさい!「自分神様	
1	2	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'『'	U+300E	4	『永平広録』「上堂語・小參」全訳注	
-	1	1	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'』'	U+300F	4	『永平広録』「上堂語・小參」全訳注	
1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'〔'	U+3010	2	【力】【フェイ】猫	
-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'〕'	U+3011	2	【力】【フェイ】猫	
-	3	-	3	1	2	-	-	1	-	-	1.	-	-	1.	-	-	-	-	-	-	-	-	-	'～'	U+301C	12	Aqours magazine~TAKAMI CHIKA~いばら	
-	1	5	1	-	3	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'ア'	U+30A1	13	ファイナンス 令和5年2月号	
-	7	1	-	2	-	-	1	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	'ヴ'	U+30F4	12	篠崎ヴァイオリン教本 1!ソヴィエト超	
-	70	10	32	12	25	11	4	3	4	2	1	2	3	-	-	1	-	-	-	-	-	-	-	'..'	U+30FB	180	熟語・ことわざ・なりたちで覚える漢字	
-	107	61	46	33	28	23	22	10	6	6	5	7	-	-	1.0	-	2	-	-	-	-	-	-	'-'	U+30FC	357	ロード・エルメロイII世の事件簿 10!二	
9	13	4	4	5	4	2	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'-'	U+4E00	46	一般常識と時事問題をひとつひとつわが		
-	128	101	238	169	193	167	136	136	79	56	44	35	19	21	11	12	5	5	1.	3	1	4	1.0	2	\$	end	1567	安岡章太郎短篇集 あくたれラルフがっ

- オプションの-eでいくつもの正規表現を指定して1文字と見なせるようにした。
  - Unicodeの「プロパティエスケープ」を用いた。Perlの正規表現の¥pを使っている。(¥はバックスラッシュ)
    - プロパティエスケープは、10年程度以内で変遷があるので注意。¥p{Sc-Hiragana}とすべしかも。
- 使われている文字が把握できる。
  - ローマ字やハート型の形状の文字など。
  - '...' (U+2026) などが文字コードと共に分かって便利。

# 利用例2

新型コロナウィルス感染事例データ（FASTALERT/JX通信社）

<https://fastalert.jp/realtme-api>

# データの概要

```
> colsummary --jg2 jx.tsv
列番 異なる値 数値化平均 列名 値の範囲 最頻値 頻度(重複) 行数
1 32563 23676.127 id 012~45836 40405|42128 1(32563) 1~5
2 30380 0.028 name1 清瀬中里郵便局~Y.笹塚ビル 1F イオンレイクタウンmori 1F|mozoワンダーシティ 1F 17(2)|14~2(722)|1(29231) 2~52
3 8307 NaN name2 |#602 カフェ&ダイナー~M Dビジネスパートナー |化粧品売場 23969|9~2(166)|1(8100) 0|1~50
4 27555 0.247 address 1F~3F, そごう大宮店, パーキング館, 1丁目-8-4 桜木町 大宮区 さいたま市 埼玉県 330-0854 日本~6 丁目-1-55 上本町天王寺区 大阪市 大阪府 543-0001 日本 東京都新宿区新宿3丁目38-1 |埼玉県越谷市レイクタウン3丁目1-1 61|40~2(1207)|1(25553) 6~69
5 32208 35.637 lat 24.3337133~45.44063102 34.7059258|33.5443949 8|6~2(283)|1(31894) 5~11
6 32157 137.628 long 124.1204775~145.7491303 135.5643221|130.3996347 8|6~2(323)|1(31798) 6~11
7 146 3.115 reports |1~494 1|2 25342|2721~2(24)|1(35) 0|1~3
8 2 0.000 disinfected false~true truelfalse 17299|15264 4~5
9 553 2020.632 first_release1 2020-01-20~2021-08-31 2021-08-24|2021-08-18 360|353~2(16)|1(22) 10
10 2 0.000 first_release2 false~true false|true 31889|674 4~5
11 22474 2020.857 last_edited 2020-04-22 15:00:00.000000 UTC~2021-12-27 11:53:00.000000 UTC 2021-01-15 13:51:00.000000 UTC|2021-01-15 14:16:00.000000 UTC 18|16~2(3097)|1(17359) 30
32,564 line(s) read; 1.675501 seconds (colsummary)
>
```

- 1列目 **id** は、連続しない自然数で、全て異なる。
- 3列目 **name2** は、空文字列が多い。
- 4列目 **address** は大部分が通常の住所と思われる。
- 8列目と10列目は、共に、**true** または **false** のみ。
- 5, 6, 7, 9, 11 列目は数値が多いが、期待されるかどうか、気になるところである。  
↑ **digitdemog** で解決。

# 住所文字列の揺らぎが分かる

- -e でいろんな正規表現パターンでまとめた。
  - アルファベットやテや数字やカタカナで始まる文字列は、英語のような住所の書き方(日本語と逆順)になっていた。
  - 都道府県が無かったり「日本」を含んでいたりする。
  - 今までの他のやり方よりも、様々な住所の(あまり正しくない)書き方の例がよく分かる。
    - 訂正するための書式やアルゴリズムの提案が可能になると考えられる。

App::digitdemog@0.073 もしくは0.073以前で作成(古い)。

- ・緯度・経度・日付・日時として、各桁の並びに異常は認められない。

# 利用例3

国立国会図書館デジタルコレクション書誌情報

<https://www.ndl.go.jp/jp/dlib/standards/opendataset/>

主に「古典籍」について

# 出版年の書式について

- いろんな(統一性の無いが便宜性のある)記号の使い方を発見。
    - 括弧も多様
    - 伏せ字のようなものも多様。
  - 改行文字を除いた文字列長としての最大値は、17文字。

# 出版年(W3CDTF)の書式について

```
> cut -f9 dataset_202207_k_internet.tsv | sed 1q
出版年(W3CDTF)
> cut -f9 dataset_202207_k_internet.tsv | digitdemog -= -g"; 3" -n0 -o1 | expandtab
 1   2   3   4   5   6   7   8   9   10  11   char  code freq example; ..; example
 0   0   0   1.  0   0   0   0   0   0   0   ' '  U+20 1   740
 0   0   0   137. 0   0   0   0   0   0   0   '-'  U+2D 137  1847-18
 9   7372  15962 16444 0   0   0   141 208 0   '0'  U+30 40136 0871; 0700; 0770; 1875||1880; 1847||1850; 1808||1810
46565 20   3283  3210 0   137. 1780. 3   71   101 0   '1'  U+31 55170 1852; 1853; 1855; 1868||1871; 1830||1831; 1906||1911
 0   68   2946  3335 0   0   0   21   96   433 0   '2'  U+32 6899  1283; 1288; 1248; 1809||1822; 1862||1862; 1895||1902
 0   373   3241  3889 0   0   0   68   316  84 0   '3'  U+33 7971  1354; 1300; 1332; 1838||1853; 1898||1903; 1911||1913
 0   175   3496  2877 0   0   0   18   370  137 0   '4'  U+34 7073  1400; 1495; 1413; 1339||1374; 1812||1864; 1852||1854
 0   1394  4842  2938 0   0   0   36   177  254 0   '5'  U+35 9641  1597; 1554; 1516; 1853||1855; 1862||1865; 1854||1855
 0   8172  4546  3424 0   0   0   1    162  170 0   '6'  U+36 16475  1649; 1600; 1668; 1895||1896; 1893||1896; 1853||1856
 1.  6288  2663  3672 0   0   0   21   198  125 0   '7'  U+37 12968  740 ; 1796; 1740; 1776||1777; 1811||1867; 1852||1857
 0   20977 2458  3105 0   0   137. 1343 52   90 0   '8'  U+38 28162  1852; 1853; 1855; 1836||1838; 1856||1858; 1881||1928
 0   1736  3138  3680 0   0   0   269  197  178 0   '9'  U+39 9198  1901; 1907; 1937; 1508||1509; 1597||1599; 1796||1799
 0   0   0   1780. 1780. 0   0   0   0   0   0   '1'  U+7C 3560  1786||1853; 1808||1809; 1876||1892; 1881||1928; 1852||1854; 1796||1799
35531 0   0   0   44658 0   0   137. 0   0   1780. $   end  82106 ; 1852; 1853; 1881||1928; 1852||1854; 1796||1799
> █
```

- 年は4桁だが3桁が、8万21016件中1個あり。
  - "740"とあり、最後に空白文字がある。
  - 西暦999年以前は、最初の桁は0で始まり、9件。
- 5と6桁目が縦線2本のものあり。
  - ORの記号かと思いきや、隣り合った2年でも無い。
- ハイフンを使った"1847-18"が137件。

# 「卷次」について

```

> cut -f3 dataset_202207_k_internet.tsv | sed 1q
巻次
> cut -f3 dataset_202207_k_internet.tsv | digitdemog --= -e'[\p{Han}あ-んア-ン]+'-e'[0-9]+'-e'[A-Za-z]+'-e'[0-9]+'-g"1;" -0- -n0 | csel -p-4..-1
char code freq example; ..: example
[ \p{Han}あ-んア-ン]+ --- 34648 第1帖; [2] 南北品川町 本芝町 下高輪村 芝田町 三田町 芝金杉町 増上寺門前 芝西応寺門前 愛宕下 天徳寺門前 飯倉町 麻布谷町 麻布新町 麻布長坂町 麻布坂下町 麻布
南日ヶ窪 白銀台町 永峯町 上大崎六軒茶屋 中目黒村 中渋谷新町 四谷塙町 今井町 下一木赤坂伝馬町 上一木駮ヶ橋 原宿村 角筈村 四谷伝馬町 内藤宿 麴町 市谷 南八町堀 南茅場町 東湊町 本湊町
船松町 南新堀 霊岸鷲白銀町 霊岸鷲長崎町 霊岸鷲浜町 南小田原町 木挽町
[0-9]+ --- 1013 1 伊(いへと); [168] 三分冊ノ一 「最樹院様七回御忌法事」 東叡山御赦帳写(天保四年)、東叡山赦帳之内書抜(天保4年)
[A-Za-z]+ --- 17 Vol. 2; [96] 木内芳軒書簡(大沼枕山宛 11月7日 後半(前半はWB32-3[6])) 
[0-9]+ --- 54913 1. d; 目録,卷109,110-114,160-162,197-200,207-209,236-239,262-265,488,517,521,523,527,533,611,629,631,639,644-645,653,673-676,682-683,709-710,712,719-720,728-730
' U+20 8680 一号 分冊の1(文化9); 卷2-30, 51, 53-62, 67, 69, 70, 81, 82, 84, 95下
' ( U+28 767 (表紙); [96] 木内芳軒書簡(大沼枕山宛 11月7日 後半(前半はWB32-3[6])) 
' U+29 762 (表紙); [96] 木内芳軒書簡(大沼枕山宛 11月7日 後半(前半はWB32-3[6])) 
' U+2C 634 中, 下巻; 目録,卷100,109,110-114,160-162,197-200,207-209,236-239,262-265,488,517,521,523,527,533,611,629,631,639,644-645,653,673-676,682-683,709-710,712,719-720,728-730
' U+2D 2623 初-4編; 目録,卷100,109,110-114,160-162,197-200,207-209,236-239,262-265,488,517,521,523,527,533,611,629,631,639,644-645,653,673-676,682-683,709-710,712,719-720,728-730
' U+2E 19 1. d.; 第2冊 略解 1巻、巻名。巻1-2
' U+2F 15 [12] 新類(第二輯) 七 賄賂之部 賄賂謝礼金等差出候類・賄賂謝礼金等之所持いたし候類 /博奕之部 博奕いたし候類・賭事いたし候類・博奕又者賭事之宿いたし候類・博奕又者賭事之世話いたし候類 /附火之部 盗可致ため附火いたし者或火を用盜入候類・遺恨ニ而附火いたし候類・怪火取扱候もの
' U+3A 2 第1冊: 上; 第2冊: 下
[ U+5B 41353 [1]; 卷20,21,36,39,49,[55-57],58-59,68,72,[76],79,[81]
] U+5D 41353 [1]; 卷20,21,36,39,49,[55-57],58-59,68,72,[76],79,[81]
' U+2015 11 [2] -様-回-御忌法事之御赦ニ付前々伺無之博奕一件三而御仕置ニ成候者共書付; [25] 向方御赦例書 御家人之部 上 御家人之部-養子一件、高利、御仕置不用もの、口論、不念、父之科、雜、理不尽、質直主、我意、直訴并箱訴、等閑取計、追放者世話、遺恨
'* U+203B 4 第95冊分冊ノ1 有徳院様一回御法事御赦帳(宝曆三年) ***三回忌; [160] 町会所一件書留目録 一 寛政四子年~十二申年分の目録 *頭に、802-028「町法改正積金起立書(三冊)」の目録あり
■ U+25A0 2 [24] 甲類(第一輯) 十 下 巧事取扱之部 居催促又者不筋之催促等いたし候類・不実商井世話いたし候類・金子押借之取持又者使等いたし候類・追放も等を罷置又者尋之ものを乍存其儘ニいたし置或者吟味筋之ものを為立退候類・人勾引之類・余人之不埒を引受又者かばひ候類・相役取計ニ泥ミ諸取書印形相違いたし候儀乍心得諸取候もの・■海鼠抜賣いたし候を差押候節身分を偽候得共取押候功を以御咎無之もの; [25] 甲類(第一輯) 十一 上 盗賊之部 御場所柄ニ而盗いたし又者御用之品等を盜候類・人を殺又者痴附盜いたし者或病入死人等之品盜候類・メリ有之長持■籠箱等を固辞明又者封状切解候類
□ U+25A1 10 島左近書状(残簡) (口)月十七日; [8] 五 慶応二年 内表紙に「實取箇綴 慶応二年 給口場」とあり
● U+25CF 1 恒賢・忠・山公・大中臣時広
' U+3000 5204 後編 一; [2] 南北品川町 本芝町 下高輪村 芝田町 三田町 芝金杉町 増上寺門前 芝西応寺門前 愛宕下 天徳寺門前 飯倉町 麻布谷町 麻布新町 麻布長坂町 麻布坂下町 麻布南日ヶ窪
白銀台町 永峯町 上大崎六軒茶屋 中目黒村 中渋谷新町 四谷塙町 今井町 下一木赤坂伝馬町 上一木駮ヶ橋 原宿村 角筈村 四谷伝馬町 内藤宿 麴町 市谷 南八町堀 南茅場町 東湊町 本湊町 船松町 南新堀 霊岸鷲白銀町 霊岸鷲長崎町 霊岸鷲浜町 霊岸鷲塙町 南小田原町 木挽町
' U+3090 8 の みや; 第10冊 (さゝれいし)
' U+30F5 3 [130] 三分冊ノ三 (御転任御任槐御祝儀・若君様御弘御祝儀・御昇進御位階御祝儀御赦ニ付手限御仕置相成候者共書付 (六カ)); [129] 三分冊ノ二 (御転任御任槐御祝儀・若君様御弘御祝儀・御昇進御位階御祝儀御赦ニ付手限御仕置相成候者共書付 (六カ))
' U+30F6 70 内藤新宿千駄ヶ谷絵図; [2] 南北品川町 本芝町 下高輪村 芝田町 三田町 芝金杉町 増上寺門前 芝西応寺門前 愛宕下 天徳寺門前 飯倉町 麻布谷町 麻布新町 麻布長坂町 麻布坂下町 麻布
南日ヶ窪 白銀台町 永峯町 上大崎六軒茶屋 中目黒村 中渋谷新町 四谷塙町 今井町 下一木赤坂伝馬町 上一木駮ヶ橋 原宿村 角筈村 四谷伝馬町 内藤宿 麴町 市谷 南八町堀 南茅場町 東湊町 本湊町 船松町 南新堀 霊岸鷲白銀町 霊岸鷲長崎町 霊岸鷲浜町 霊岸鷲塙町 南小田原町 木挽町
' U+30FC 5 [6] 四 分冊ノ一 天保三一年四; [12] 六 三分冊ノ三 天保六一七年
' ( U+FF08 1307 (帙); 3(2) 古金銀引替高(天保13年5月)・根津門前模毒院取建調(明治元年2月)・鈴木常太郎等原市之進ヲ殺害一件(慶応3年8月)・櫻田一件書抜(文久元年7月)・大橋順藏一件書抜(万延2年閏8月)・飯泉喜内一件書抜(文久2年11月)
' ) U+FF09 1311 (帙); 3(2) 古金銀引替高(天保13年5月)・根津門前模毒院取建調(明治元年2月)・鈴木常太郎等原市之進ヲ殺害一件(慶応3年8月)・櫻田一件書抜(文久元年7月)・大橋順藏一件書抜(万延2年閏8月)・飯泉喜内一件書抜(文久2年11月)
' U+FF0C 3 卷第1, 2; 卷第6, 7
' U+FF0D 30 卷1 - 3; [160] 町会所一件書留目録 一 寛政四子年~十二申年分の目録 *頭に、802-028「町法改正積金起立書(三冊)」の目録あり
' [ U+FF3B 4 [25] 十一 分冊ノ一 [破損] 御祝儀之御赦ニ付 [破損] 成候者共書付; [26] 十一 分冊ノ二 ([破損] 御祝儀之御赦ニ付 [破損] 成候者共書付)
' ] U+FF3D 4 [25] 十一 分冊ノ一 [破損] 御祝儀之御赦ニ付 [破損] 成候者共書付; [26] 十一 分冊ノ二 ([破損] 御祝儀之御赦ニ付 [破損] 成候者共書付)
' ~' U+FF5E 53 卷1~2; [165] 町会所一件書留目録 六 文政十三(天保元)年~天保八酉年分の目録
' -' U+FF86 2 第16冊分冊ノ1 家治公 日光山御社參御赦ニ付前々御仕置相成候者共書付 一; 第16冊分冊ノ2 (家治公 日光山御社參御赦ニ付前々御仕置相成候者共書付 一)
' ' U+FF89 1 [8] 竪川之部三 分冊ノ二 文化5年
$ end 82106 ; [2] 南北品川町 本芝町 下高輪村 芝田町 三田町 芝金杉町 増上寺門前 芝西応寺門前 愛宕下 天徳寺門前 飯倉町 麻布谷町 麻布新町 麻布長坂町 麻布坂下町 麻布南日ヶ窪 白銀台町 永峯町 上大崎六軒茶屋 中目黒村 中渋谷新町 四谷塙町 今井町 下一木赤坂伝馬町 上一木駮ヶ橋 原宿村 角筈村 四谷伝馬町 内藤宿 麴町 市谷 南八町堀 南茅場町 東湊町 本湊町 船松町 南新堀 霊岸鷲白銀町 霊岸鷲長崎町 霊岸鷲浜町 霊岸鷲塙町 南小田原町 木挽町
>

```

- 使われた文字列や文字が分かる。
- コマンド`csel`の機能で最も右の4列だけ選んだ。

# 「シリーズ」について

```
> cut -f4 dataset_202207_k_internet.tsv | sed 1q
シリーズ
> cut -f4 dataset_202207_k_internet.tsv | digitdemog -= -e'[\p{Han}]+'-e'[あ-んア-ン]+'-e'[0-9]+'-e'[A-Za-z]+'-g"2;" "-0- -n0
0 1 2 3 4 5 6 7 8 9 10 char code freq example; ..; example
10570 218 784 230 257 7 107 219 34. - - - [\p{Han}]+ --- 12426 春見立当羽子板；花鳥錦絵；田中叢書；第1-3冊；杉山叢書；第35-36冊
147 908 15 90 12 64 - - 55. - - - [あ-んア-ン]+ --- 1291 つきの百姿；いこくことば；傾城道中双（ロク） 見立よしはら五十三つい
- - - - - - - - - - - [0-9]+ --- 0
- - - - - - - - - - - [A-Za-z]+ --- 0
- 19 - 1151 462 550 320 122 - 6. - - [0-9]+ --- 2630 築喜廬叢書之2；津逮祕書第9集；演劇台帳。15-17,261,262；〔幸若舞集〕；[25]
- 328. 1374. 395 - 67. - - - - - U+20 2164 錦窯翁遺書；第1冊；錦窯翁遺書；第37冊；〔幸若舞集〕；[9]；〔幸若舞集〕；[25]
- - - - - - - - - - - ;' U+2C 53 書記群類。7,8,9；演劇台帳。226,227,325；演劇台帳。15-17,261,262
- - - - 412 305 34. - - - - - U+2D 751 演劇台帳。3-7；演劇台帳。15-17,261,262；田中叢書；第1-3冊；杉山叢書；第35-36冊
- 1374. - - - - - - - - - - - U+2E 1374 古曆集。20；演劇台帳。1；松島松作芝居台帳。3；演劇台帳。162-165
- - 328. - 67. - - - - - ;' U+3B 395 錦窯翁遺書；第1冊；錦窯翁遺書；第37冊；〔幸若舞集〕；[9]；〔幸若舞集〕；[25]
82. - - - 4. - 66. - - - - - '[' U+5B 152 [江戸切絵図]；〔幸若舞集〕；[5]；〔幸若舞集〕；[9]；〔幸若舞集〕；[25]
- - 82. - - - 4. - 66. - - - - - ']' U+5D 152 [江戸切絵図]；〔幸若舞集〕；[5]；〔幸若舞集〕；[9]；〔幸若舞集〕；[25]
2 70 - 55. - - - - - - - - - U+3000 127 ；濡衣女鳴神 近江八勇之内；傾城道中双（ロク） 見立よしはら五十三つい；観音靈験記 秩父巡礼
- - - 5 - - - - - - - 'ゞ' U+309E 5 てい女みさほかゞみ
- 5 - - - - - - - - - - 'ヶ' U+30F6 5 五ヶ国人物；美人十二ヶ月
2. - - - - - - - - - - - '(' U+FF08 2 (無題)
- - 2. - - - - - - - - - - ')' U+FF09 2 (無題)
71303 7881 337 659 701 217 456 156 290 100 6. $ end 82106 ; 春見立当羽子板；演劇台帳。15-17,261,262；〔幸若舞集〕；[25]
>
```

- 使われた文字列や文字が分かる。

# 上海新華書店の「請求番号」

```

> cut -f11 xinhua_bib.txt | sed 1q
請求記号
> cut -f11 xinhua_bib.txt | digitdemog -= -L4
length freq minstr maxstr first_str last_str
6 36 'XP-A-1'(1) 'XP-D-9'(1) 'XP-A-1'(1) 'XP-D-9'(1)
7 360 'XP-A-10'(1) 'XP-D-99'(1) 'XP-A-10'(1) 'XP-D-99'(1)
8 3600 'XP-A-100'(1) 'XP-D-999'(1) 'XP-A-100'(1) 'XP-D-999'(1)
9 20326 'XP-A-1000'(1) 'XP-D-1045'(1) 'XP-A-1000'(1) 'XP-D-1045'(1)
10 116015 'XP-A-10000'(1) 'XP-B-46080'(1) 'XP-A-10000'(1) 'XP-B-46080'(1)
> cut -f11 xinhua_bib.txt | digitdemog -= -n0 -o1 -g3 | expandtab
1 2 3 4 5 6 7 8 9 10 11
0 0 140337. 0 140337. 0 0 0 0 0 0
0 0 0 0 0 14541 14081 13644 11603 0
0 0 0 0 23490 14496 14000 13643 11601 0
0 0 0 0 23444 14477 14000 13642 11600 0
0 0 0 0 22723 14395 14000 13640 11602 0
0 0 0 0 18525 14396 13995 13633 11602 0
0 0 0 0 12443 14396 13990 13633 11602 0
0 0 0 0 12444 13477 13990 13632 11602 0
0 0 0 0 12444 13396 13989 13632 11601 0
0 0 0 0 12380 13396 13979 13620 11601 0
0 0 0 0 2444 13331 13917 13622 11601 0
0 0 89933 0 0 0 0 0 0 0
0 0 46079 0 0 0 0 0 0 0
0 0 3280 0 0 0 0 0 0 0
0 0 1045 0 0 0 0 0 0 0
0 140337. 0 0 0 0 0 0 0 0 0
140337. 0 0 0 0 0 0 0 0 0
0 0 0 0 0 36 360 3600 20326 116015 $
```

	<u>char</u>	<u>code</u>	<u>freq</u>	<u>example1..example</u>
'-	U+2D	280674	XP-A-1 XP-A-2 XP-A-3 XP-D-1043 XP-D-1044 XP-D-1045	
'0'	U+30	53869	XP-A-10 XP-A-20 XP-A-30 XP-B-46060 XP-B-46070 XP-B-46080	
'1'	U+31	77230	XP-A-1 XP-A-10 XP-A-11 XP-B-46051 XP-B-46061 XP-B-46071	
'2'	U+32	77163	XP-A-2 XP-A-20 XP-A-21 XP-B-46052 XP-B-46062 XP-B-46072	
'3'	U+33	76360	XP-A-3 XP-A-30 XP-A-31 XP-B-46053 XP-B-46063 XP-B-46073	
'4'	U+34	72151	XP-A-4 XP-A-40 XP-A-41 XP-B-46054 XP-B-46064 XP-B-46074	
'5'	U+35	66064	XP-A-5 XP-A-50 XP-A-51 XP-B-46055 XP-B-46065 XP-B-46075	
'6'	U+36	65145	XP-A-6 XP-A-60 XP-A-61 XP-B-46056 XP-B-46066 XP-B-46076	
'7'	U+37	65062	XP-A-7 XP-A-70 XP-A-71 XP-B-46057 XP-B-46067 XP-B-46077	
'8'	U+38	64976	XP-A-8 XP-A-80 XP-A-81 XP-B-46058 XP-B-46068 XP-B-46078	
'9'	U+39	54915	XP-A-9 XP-A-90 XP-A-91 XP-B-46059 XP-B-46069 XP-B-46079	
'A'	U+41	89933	XP-A-1 XP-A-2 XP-A-3 XP-A-89934 XP-A-89935 XP-A-89936	
'B'	U+42	46079	XP-B-1 XP-B-2 XP-B-3 XP-B-46078 XP-B-46079 XP-B-46080	
'C'	U+43	3280	XP-C-1 XP-C-2 XP-C-3 XP-C-3278 XP-C-3279 XP-C-3280	
'D'	U+44	1045	XP-D-1 XP-D-2 XP-D-3 XP-D-1043 XP-D-1044 XP-D-1045	
'P'	U+50	140337	XP-A-1 XP-A-2 XP-A-3 XP-D-1043 XP-D-1044 XP-D-1045	
'X'	U+58	140337	XP-A-1 XP-A-2 XP-A-3 XP-D-1043 XP-D-1044 XP-D-1045	
end	140337		XP-A-1 XP-A-2 XP-A-3 XP-B-46078 XP-B-46079 XP-B-46080	

- XP- の A, B, C, D とあって、1から999までは少なくとも連番。
- XP-Bに付いては1から46080あるが1箇所で抜けがありそう。
  - 調べてみたら、8078が抜けていた。
  - XP-Aにおいても同様そう。それは表のA,B,C,Dを見て分かった。

# 上海新華書店の「タイトル」

```
> cut -f1 xinhua_bib.txt | digitdemog -.6 -= -g1◆ -0- -n0 --width 6 -e'[0-9]+' -e'[a-zA-Z]+' -e'[[:^ascii:]]+' -y100.. | expandtab | less -XRS
140,338 line(s) read; "タイトル" is the 1st line; 2.071973 seconds (digitdemog)
0   1   2   3   4   5   char      code freq example◆...◆example
484 505 282 493 53 44 [0-9]+    --- 1861 1950年的音乐运动 ◆A50-400 1500型压缩机的制造、安装、使用
240 159 46 50 20 14 [a-zA-Z]+    --- 529 CC豪門资本内幕 ◆分光化学 ; 应用X射线化学
139607 553 2976 1486 2261 167 [[:^ascii:]]+ --- 147050 我们的塘沽新港 ◆上海-50型轮式拖拉机零件图册 第2版
- 3442 592 1503 118 1055 ' ' U+20 6710 识小录 再版 ◆苕溪鱼隐丛话 ; 说诗啐语 ; 佩文诗韵释要
- 52   - 4   76   -  '!' U+21 132 用毛泽东思想武装起来、为争取文艺的更大丰收而奋斗!◆人啊，人！
- 544  4   535 16   31  ' "' U+22 1130 鲁迅"野草"探索 ◆批判"四人帮"对"唯生产力论"的"批判"
-   - 259 2.   6   5   '(' U+28 272 银海指南 (眼科大成)◆1956-1967年科学技术发展远景规划纲要 (修正草案) 通俗讲话
-   -   - 251 2.   ')' U+29 253 银海指南 (眼科大成)◆1958年度国营工业企业基本业务标准定期 (月、季) 会计报表格式和说明
- 388  4   4   48  3   ',' U+2C 447 国营企业经理, 厂(矿)长国家统考文件, 复习大纲汇编 ◆社会主义企业组织与生产管理的原理 ; 材
- 673 442 21   26  15  '-' U+2D 1177 黎培里-帝国主义的侵略工具 ◆中华人民共和国法规汇编1960年1月-6月
- 172 33   74  2   36  ':' U+3A 317 国学汇编之八: 史学纂要 ◆中国近代史资料丛刊 第33种: 中法战争
-   - 1323 12   6   6   ';' U+3B 1347 救伤秘旨 ; 跌损妙方 ◆醒来的"睡神" ; 奇猎记 ; 苏平凡原著 ; 胡翀改编 ; 叶坚画绘
- 192   1   8   12  5   '?' U+3F 218 文字改革和汉字简化是怎么回事?◆什么是"人民之友"以及他们如何攻击社会民主主义者?
- 133605 746 1779 1292 1524 $ end 138946 我们的塘沽新港 ◆越剧西厢记舞台服装设计 ; 梁山伯与祝英台舞台服装设计
(END)
```

# 上海新華書店の「シリーズ」

series																code	freq	example	example	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	char	code	freq	example	example
14963	54	54	21	2439	440	49	3.	0	1760.	0	0	0	0	0	0	\p{Han}+	---	19783	史地小丛书 萬有文庫；第二集；360自然科學小叢書	
0	40	0	0	1.	0	1.	0	1.	0	1.	0	1.	0	1.	0	[A-Za-z]+	---	46	中學學科自測ABC 中美關係研究叢書 = Series studies in Sino-American relations	
58	10	8	0	3446	34	83	0	1783	0	0	0	0	0	0	0	[0-9]+	---	5422	1958年全國農業展覽會 萬有文庫；第二集；360自然科學小叢書	
0	5894	1	5879	0	1857	0	1784	0	1.	0	0	0	1.	0	0	'.'	U+20	15417	共產黨知識叢書(共六冊) 中美關係研究叢書 = Series studies in Sino-American relations	
0	5	0	1	0	4.	0	0	0	0	0	0	0	0	0	0	'''	U+22	10	慶祝“六一”國際兒童節特輯 高考指導“3+2”叢書	
0	0	16.	0	0	0	0	0	0	0	0	0	0	0	0	0	'('	U+28	16	共產黨知識叢書(共六冊) 华东華中區高等林學院(校)教學用書	
0	0	0	16.	0	0	0	0	0	0	0	0	0	0	0	0	')'	U+29	16	共產黨知識叢書(共六冊) 华东華中區高等林學院(校)教學用書	
0	0	0	4.	0	0	0	0	0	0	0	0	0	0	0	0	'+'	U+2B	4	高考指導“3+2”叢書	
0	4	0	0	0	4	0	0	0	0	0	0	1.	0	0	0	'-'	U+2D	9	1958-1959學年度上海市高中畢業班複習參考資料 中美關係研究叢書 = Series studies in Sino-American relations	
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	'.'	U+2E	1	農業生產技術基本知識. 第20分冊	
0	0	5876	0	1	0	1765	0	0	0	0	0	0	0	0	0	';;'	U+3B	7642	廣西青年詩丛；含羞草 萬有文庫；第二集；360自然科學小叢書	
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	'<'	U+3C	1	萬有文庫<>簡編漢譯世界名著	
0	0	1.	0	0	0	0	0	0	0	0	0	0	0	0	0	'='	U+3D	1	中美關係研究叢書 = Series studies in Sino-American relations	
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	'>'	U+3E	1	萬有文庫<>簡編漢譯世界名著	
0	41	0	2	0	0	0	0	0	0	0	0	0	0	0	0	'..'	U+B7	43	自然·人·哲學	
125316	8971	93	50	4	3564	441	111	3.	23	1760.	0	0	0	0	1.	\$	end	140337	中美關係研究叢書 = Series studies in Sino-American relations	

# その他重要事項

# 併用したコマンド

- `cSEL` : 指定位置の列を抽出。`cut` を拡張。
- `colsummary` : 全列の性質を整理して表示。
- `expandtab` : TSV形式の表を立て揃えにする。

# さらに実装したい機能

- 各頻度の値に対応する「行の集合」の包含関係を検出する。.1 や .2などで区別をつけて表示。
  - ただし頻度1に対して、多数の包含関係の親が発生したり、分岐が発生したりして厄介。
  - 重要度の高い、頻度の高い値に対して、優先的に実行したい。
- Exampleの区切りの|を変更可能にしたい。
  - Exampleも|で切ったことが分かるようにしたい。
- -g3. でなくて-g3Rのようにしたい。
  - -g3. と指定したら ピリオド(.)で例を区切るように・。

# 気になること

- 世界標準時として記録されていた。
  - 時間の分布から、設定ミスか正しいか分かりそう。
- 各桁の並びを別個に見ただけ。
  - 本当に正しく日付または日時を表しているの検査が必要。
  - さらに自作のプログラムが必要! datecheckとか。
- 機能を増やしたことによる動作速度の低下