

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(uC_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。

利用例

`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で保存も出来て
使い勝手も良い。

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/
- 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) 行数
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 8|1|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

cut -f1 TRCOpenBibData_20230225.txt digitdemog -.0																		char	code	freq	example..lexample			
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	'.'	U+2D	6252	978-4-00-027249-0\n 978-4-9911149-4-6\n		
0	0	0	1563	0	1563	0	0	477	493	358	181	52	2	0	1563	0	0	'0'	U+30	1817	978-4-00-027249-0\n 978-4-09-872013-2\n			
0	0	0	0	0	0	392	140	88	93	242	215	144	196	149	0	158	0	0	'1'	U+31	1470	978-4-10-603071-0\n 978-4-19-901093-4\n		
0	0	0	0	0	0	85	135	86	135	206	226	134	152	168	0	143	0	0	'2'	U+32	1437	978-4-251-07862-9\n 978-4-299-04076-3\n		
0	0	0	0	0	0	106	109	80	107	140	204	175	185	178	0	153	0	0	'3'	U+33	1448	978-4-300-10114-8\n 978-4-398-29674-0\n		
0	0	0	0	0	0	121	117	119	95	196	191	170	134	158	0	147	0	0	'4'	U+34	3026	978-4-00-027249-0\n 978-4-9911149-4-6\n		
0	0	0	0	0	0	1563	0	97	263	151	75	65	142	199	141	168	0	162	0	0	'5'	U+35	1279	978-4-502-45271-0\n 978-4-596-76929-9\n
0	0	0	0	0	0	151	95	136	130	52	119	147	166	148	0	135	0	0	'6'	U+36	1331	978-4-620-79462-4\n 978-4-651-20312-6\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\n 978-4-9911149-4-6\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\n 978-4-9911149-4-6\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\n 978-4-9911149-4-6\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	1563	\$	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桁目に掛けて、数字の0~9が現れる。

- ハイフンが現れることがある。
- 数値の分布は、各桁において、大体一様なことあれば、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:

<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>example1.. example</u>
-	143	-	-	6	-	-	-	'0'	U+30	149	20cm 20×20cm
656	304	-	1	5	-	-	-	'1'	U+31	966	18×26cm 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 24×26cm
-	139	-	-	1	-	-	-	'5'	U+35	140	15cm
-	204	-	-	20	-	-	-	'6'	U+36	224	26cm
-	29	-	-	2	-	-	-	'7'	U+37	31	27cm 17×22cm
-	91	-	-	-	-	-	-	'8'	U+38	91	18×26cm 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	18×26cm 17×22cm
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

```
> digitdemog <(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}'  
0 1 2 3 4 5 6 7 8 9 10 char code freq example..example  
0 0 0 0 0 0 0 0 0 0 \d{8} --- 0 978-4-9911149-3-9\n|978-4-9911149-4-6\n  
0 0 0 0 2 0 0 0 0 0 \d{7} --- 2 978-4-901794-26-8\n|978-4-944185-19-1\n  
0 0 0 0 52. 0 477. 0 0 0 \d{6} --- 529 978-4-8053-144-7\n|978-4-89831-969-7\n  
0 0 0 0 181. 0 493. 0 0 0 \d{5} --- 674 978-4-7503-5527-6\n|978-4-8471-4938-2\n  
0 0 0 0 358. 0 358. 0 0 0 \d{4} --- 716 978-4-00-027249-0\n|978-4-9911149-4-6\n  
1563. 0 0 0 493. 0 181. 0 0 0 \d{3} --- 2237 978-4-00-027249-0\n|978-4-19-901093-4\n  
0 0 0 0 477. 0 52. 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 '-' U+2D 6252 978-4-00-027249-0\n|978-4-9911149-4-6\n  
0 0 0 0 0 0 0 158 0 0 '0' U+30 158 978-4-00-027249-0\n|978-4-910815-06-0\n  
0 0 0 0 0 0 0 143 0 0 '1' U+31 143 978-4-01-093297-1\n|978-4-944185-19-1\n  
0 0 0 0 0 0 0 153 0 0 '2' U+32 153 978-4-00-602349-2\n|978-4-910539-07-2\n  
0 0 0 0 0 1 0 147 0 0 '3' U+33 148 978-4-9911149-3-9\n  
0 0 1563. 0 0 0 1 0 162 0 0 '4' U+34 1726 978-4-00-027249-0\n|978-4-9911149-4-6\n  
0 0 0 0 0 0 0 135 0 0 '5' U+35 135 978-4-00-312281-5\n|978-4-910803-13-5\n  
0 0 0 0 0 0 0 181 0 0 '6' U+36 181 978-4-01-035054-6\n|978-4-9911149-4-6\n  
0 0 0 0 0 0 0 148 0 0 '7' U+37 148 978-4-01-093729-7\n|978-4-910603-11-7\n  
0 0 0 0 0 0 0 162 0 0 '8' U+38 162 978-4-00-372510-8\n|978-4-924523-38-8\n  
0 0 0 0 0 0 0 174 0 0 '9' U+39 174 978-4-01-035053-9\n|978-4-9911149-3-9\n4. 0 0 0 0 0 0 0 0 1563. 0 "\n" U+0A 1567 \n  
0 4. 0 0 0 0 0 0 0 1563. $ end 1567 \n
```

- ISBNの列に対して「パターン」をいくつか手入力した。
 - 数値のまとめを、「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始まりのまとめ)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)を確かめることができた。
- **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**。₁₂

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, 524p15, 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, 524p15, 78p	
-	137	119	1	1.00000	2	4	3	-	-	-	-	-	'0'	U+30	267	100p102p	
532	136	167	1	4	18	9	7	2	13	-	-	-	'1'	U+31	889	175p143p	
445	167	93	4	8	8	6	1.00	1	-	11	-	-	'2'	U+32	744	228p1255p	
233	116	139	3	6	8	4	-	1.0	-	1.00	-	-	'3'	U+33	511	336p1387p	
84	111	101	-	2	5	2	2	1	1.000	-	-	-	'4'	U+34	309	431p1474p	
36	183	127	3	2	8	8	1	-	-	-	-	-	'5'	U+35	368	513p1526, 82p	
18	119	89	-	1	7	5	1.0	5	1.0000	1	1.	-	'6'	U+36	248	68p163p	
26	149	132	2	1.0	3	7	1	-	-	1.	-	-	'7'	U+37	322	79p177p	
21	128	101	-	1	5	4	3	1.000	-	-	-	-	'8'	U+38	264	80p182p	
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	p93~2321p97~192	
-	-	-	-	-	-	-	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16p1541p 図版16p	
-	-	-	-	1.000	-	-	-	-	-	-	-	-	'枚'	U+679A	1	23p 枚25~84	
-	-	-	-	-	-	6.	-	-	-	-	-	-	'版'	U+7248	6	292p 図版16p1541p 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|1541p 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|1526, 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|1541p 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:]+ -.2l 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) 嘸瘍		
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め		
4	44	11	36	25	20	11	10	6	6	2	4	2	2	4	-	1	-	1.0	-	-	-	-	-	[[:ascii:]]+	---	189	#シンFIRE論 <ハーフ>物語		
-	3	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'x'	U+D7	5	女子高生×日常に蠢く怪異×伸縮素材×口		
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'é'	U+E9	1	écriture新人作家・杉浦李奈の推論 8		
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'“'	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	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'…'	U+2026	2	挾啓...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、おとなちょ。		
-	1	3	2	3	3	4	4	3	2	2	1	1	1	-	-	1	1	-	-	-	-	-	-	'. '	U+3002	32	U+3005	4	威風堂々悪女 11
-	1	-	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'々'	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・小参」全訳注		
-	1	1	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'』'	U+300F	4	『永平広録』「上堂語・小参」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~ ぱら		
-	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 ニ		
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かる			
-	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年程度以内に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

JX通信社のデータ

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 1Flomo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 2396919~2(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 true|false 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

住所文字列の揺らぎが分かる																															
> cut -f4 jx.tsvl digitdemog -e '[a-z]+' -e '[A-Z]+' -e '[0-9]+' -e '[\u{00a5}-\u{00a6}]+*' -e '[\u{00a1}-\u{00a3}]+*' -e '[あ-ん]+*' -e '[\u{00a5}:\u{00a5}]+*' -g1 expandtab less -RSX																															
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	25	26	27	char	code	freq	example1...example
-	-	1	-	-	-	-	1	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	[a-z]+	---	4	愛知県名古屋市西区二方町40m o z o ワンダーシティ1F\n	
-	-	2	1	-	6	9	10	5	4	9	4	4	7	2	3	2	-	1	-	-	-	-	-	-	-	[A-Z]+	---	69	\u{00a5}1F 大阪府泉南市りんくう南浜3-12\nl神奈川県1F高津区溝口		
4	28551	293	28061	346	19488	336	2503	86	139	32	14	10	4	9	2	1	-	-	-	-	-	-	-	-	-	[0-9]+	---	79879	34-1 信達牧野 泉南市 大阪府 590-0522 日本\nl2丁目-39-1 みどり		
32522	31	22014	343	3029	156	291	190	142	77	66	32	53	20	13	8	9	5	4	3	2	2	-	1.	-	1.	[丁-龍]+	---	59014	東京都日野市万願寺1丁目18-1\nl大分県大分市宮崎760\n		
3	978	43	97	17	58	45	155	88	103	50	42	25	7	11	1	8	1	4	-	2	-	1	-	1.	-	[ア-ン]+	---	1740	イオンタウン仙台泉大沢2f, 1丁目-5-1 大沢 泉区 仙台市 宮城県		
1	1876	7	46	15	17	5	25	4	5	5	6	1	2	-	1	-	-	-	-	-	-	-	-	-	[あ-ん]+	---	2017	さいたま市緑区美園5丁目50番地1\n			
30	592	3211	574	8645	460	17027	444	2496	220	204	114	42	73	32	25	11	12	6	7	3.	2.	2	-	1.	1.	[[\u{00a5}:\u{00a5}]+]	---	34235	5Chome-45 Nishiyamamachi, Hekinan, Aichi 447-0064 日本\nl1F-		
-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'x'	U+D7	1	神奈川県座間市相模が丘1丁目12-2 コジマxピックカメラ座間店\n			
-	-	13	-	6	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'.'	U+2010	21	佐賀県小城市芦刈町芦溝三本松130-1\nl神奈川県横須賀市不入斗町			
-	-	2	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'.'	U+2015	4	岡山県岡山市東区西大寺中野677-1\nl神奈川県横浜市戸塚区品濃町554			
-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'I'	U+2160	1	北海道札幌市西区琴似4条1丁目1-1 コルテナI\n			
-	1	2	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	'.'	U+3000	5	横浜市金沢区八景島 横浜・八景島シーパラダイス\n			
-	1	26	-	1	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	'.'	U+3001	30	日本、福岡県福岡市東区香椎浜3丁目12-1\n			
-	91	-	3	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'々'	U+3005	95	石川県野々市市白山町4-1\nl千葉県印旛郡酒々井町酒々井889-			
-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	''	U+300C	1	大分県大分市要町1-40「豊後にわさぎ市場」内\n			
-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	''	U+300D	1	大分県大分市要町1-40「豊後にわさぎ市場」内\n			
-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'['	U+3010	1	愛知県名古屋市西区二方町59【ジェームス名古屋西店併設】\n				
-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'']'	U+3011	1	愛知県名古屋市西区二方町59【ジェームス名古屋西店併設】\n				
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'丁'	U+3012	3	\u{00a5}4F 東京都墨田区押上1丁目1-1 施トヤード JP 131-0045 東京都				
-	-	-	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'ヴ'	U+30F4	3	神奈川県相模原市中央区相模原2丁目7-1 ヴェルデュール相模原2				
-	185	-	28	1	-	-	-	-	-	2	4	-	1	-	-	-	-	-	-	-	-	-	-	'ケ'	U+30F6	221	岐阜県羽島市竹鼻町梅ヶ枝200-2\nl福岡県春日市星見ヶ丘4丁				
-	1	2	3	9	-	8	2	2	6	3	2	3	-	-	-	1	-	-	1	-	-	-	-	'フ'	U+30FB	43	フレル・ウイズ自由が丘2f, 1丁目-6-9 自由が丘 目黒区 東京都				
-	-	63	6	50	13	39	19	61	25	40	20	12	7	1	4	-	3	-	1	-	-	-	-	''	U+30FC	365	滋賀県蒲生郡竜王町大字薬師字砂山1178-694\nl茨城県東茨城				
-	-	255	-	7	1	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	''	U+4E00	264	愛知県一宮市朝日1丁目8-9\nl愛知県一宮市大和町宮地花池高見4				
-	-	-	1	-	-	2	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	'('	U+FF08	5	奈良県橿原市四条町 医大病院玄関口(バス)\n				
-	-	-	-	1	1	-	2	-	2	-	1	-	-	-	-	-	-	-	-	-	-	-)'	U+FF09	7	奈良県橿原市四条町 医大病院玄関口(バス)\n					
-	-	6649	36	16787	242	2346	51	119	16	3	1	1	1	-	-	-	-	-	-	-	-	-	-	''	U+FF0D	26252	新潟県三条市大野畠6-18\nl愛知県岡崎市戸崎町外山38-5\n				
-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	'.'	U+FF1A	1	熊本県菊池郡菊陽町津久礼2750-2 駐車場:有\n				
-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	'-'	U+FF3F	1	東京都品川区大井1丁目2-12 JR 大井町駅 1F 橋上 中央口改					
-	1.	-	-	-	-	-	-	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	'~'	U+FF5E	5	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				
-	-	236	3120	297	8468	329	16713	391	2411	177	180	88	28	55	26	13	9	7	3	4	3.	2.	1	-	1	1.	\$	end	32563	神奈川県横浜市戸塚区川上町\nl大阪府茨木市上穂東町3-23\n	

- 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〒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)。

```
> cut -f9 jx.tsv | digitdemog -- - . -g1 | expandtab
0 1 2 3 4 5 6 7 8 9 10 11 char code freq example
0 0 0 0 0 0 0 0 0 32563. 0 "\n" U+0A 32563 2020-02-24|2021-08-26
0 0 0 0 32563. 0 0 32563. 0 0 0 0 '-' U+2D 65126 2020-02-24|2021-08-26
0 32563. 0 11992 0 27194 1082 0 8795 3401 0 0 '0' U+30 85027 2020-02-24|2021-08-26
0 0 0 20571 0 5369 4282 0 11143 3702 0 0 '1' U+31 45067 2021-01-01|2021-08-26
32563. 0 32563. 0 0 0 3060 0 10776 3128 0 0 '2' U+32 82090 2020-02-24|2021-08-26
0 0 0 0 0 0 1298 0 1849 3119 0 0 '3' U+33 6266 2020-03-02|2021-03-31
0 0 0 0 0 0 3584 0 0 3236 0 0 '4' U+34 6820 2020-04-01|2021-04-30
0 0 0 0 0 0 3782 0 0 3046 0 0 '5' U+35 6828 2020-05-01|2021-05-31
0 0 0 0 0 0 1169 0 0 3390 0 0 '6' U+36 4559 2020-06-02|2021-06-30
0 0 0 0 0 0 3841 0 0 3401 0 0 '7' U+37 7242 2020-07-01|2021-07-31
0 0 0 0 0 0 9546 0 0 3258 0 0 '8' U+38 12804 2020-08-01|2021-08-26
0 0 0 0 0 0 919 0 0 2882 0 0 '9' U+39 3801 2020-09-01|2020-09-30
0 0 0 0 0 0 0 0 0 32563. $ end 32563 2020-02-24|2021-08-26
>
```

- 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。

その他重要事項

併用したコマンド

- `cse1` : 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。