PWS CUP 2016 トランザクションデータの書式 Ver. 0.3

匿名加工コンテストで提出する匿名加工トランザクションデータは、以下の書式に従う必要があります。従っていないデータは受け付けられません。マスターデータについては別紙を参照願います。

また、再識別コンテストで与えられる匿名加工トランザクションデータは他チームが提出したデータそのものですので、以下の書式に従っています。

元（加工前）データは本書式に従っているので、参考にしてください。ただし、本書式は元データに比べて、0 サプレスの制限が緩いなど少し緩い書式になっているため、再識別コンテストでは注意してください。

全体的な書式

全体的な書式は次の通りです。

項目	書式
ファイル形式	CSV (ASCII)、二重引用符なし
改行文字	LF (¥n) または CRLF (¥r¥n)
行数	1 行以上 1,000,000 （百万）行以下
各行の文字数	78 文字以下（改行文字は含まない）

各行は 7 列（セル）で構成されます。つまり、各行にちょうど 6 つのカンマが含まれる必要があります。空行があってはいけません。最終行の改行はあってもなくても良いです。たとえば、最終行は

12680,581587,2011/12/9,12:50,22138,4.95,3<LF>[EOF]

でも

12680,581587,2011/12/9,12:50,22138,4.95,3[EOF]

でも良いです（<LF>は改行文字、[EOF]はファイルの終わりを示します）。

各列の書式

各列の書式は次の通りです。正規表現は Java での表記（http://docs.oracle.com/javase/jp/8/docs/api/java/util/regex/Pattern.html）です。

#	通称	書式パターン（正規表現）	書式の概要
1	顧客 ID	[1-9][0-9]{0,8}	9 桁以下の自然数
2	伝票 ID	[1-9][0-9]{0,8}	9 桁以下の自然数
3	年月日	<後に掲載>	2010, 2011 年の YYYY/MM/DD 形式

4	時分	([01]?[0-9]|2[0-3]):[0-5][0-9]	hh:mm 形式
5	製品 ID	[A-Z0-9]{1,9}	9 文字以下の任意の大文字英数文字列
6	単価	<後に掲載>	整数部 5 桁以下、小数部 2 桁以下の正小数
7	数量	[1-9][0-9]{0,5}	6 桁以下の自然数

上記列 3（年月日）の書式パターン（正規表現）は次の通りです。

```
(?x) #コメントモード開始
201[01]/ #年
( #月日開始
  (0?[13578]|1[02])/(0?[1-9]|[1-2][0-9]|3[01])| #大の月
  (0?[469]|11)/(0?[1-9]|[1-2][0-9]|30)| #小の月（2 月を除く）
  (0?2)/(0?[1-9]|1[0-9]|2[0-8]) #2 月
)
```

上記列 6（単価）の書式パターン（正規表現）は次の通りです。

```
(?x) #コメントモード開始
0¥.0[1-9]|0¥.[1-9][0-9]?| #範囲（0, 1）
[1-9][0-9]{0,4}(¥.[0-9]{1,2})? #1 以上
```

各列とも、値の前後に不要な空白があるだけでも不正になることに注意してください。

列 1（顧客 ID）の注意点および例

列 1 は自然数で、元データは 5 桁ですが、匿名加工データは 9 桁以下です（符号付き 32bit 整数型に収まります）。

例：

- ○ 987654321
- × 01234　　∵ 0 が先行してはいけません。

列 2（伝票 ID）の注意点および例

列 2 は自然数で、元データは 6 桁ですが、匿名加工データは 9 桁以下です（符号付き 32bit 整数型に収まります）。

例：

- ○ 987654321
- × 012345　　∵ 0 が先行してはいけません。

列 3（年月日）の注意点および例

列 3 は「年/月/日」の書式で、元データは月と日では先行の「0」がありませんが、匿名加工データでは先行の「0」が許されます。年は「2010」か「2011」のどちらかでなくてはならず、月日は実在する値でなくてはなりません。

例：

- ○ 2010/1/1
- ○ 2011/01/01
- × 2011/1/001　　∵ 月も日も 3 桁以上に 0 を先行させてはいけません。
- × 2010/2/29　　∵ 2/29 は（2010 年には）ありません。

列 4（時分）の注意点および例

列 4 は「時:分」の書式で、元データは時では先行の「0」がありませんが、匿名加工データでは先行の「0」が許されます。一方、分は必ず 2 桁です。時は 0～23、分は 0～59 でなくてはなりません。

例：

- ○ 0:00
- ○ 00:00
- × 0:0　　∵ 分は 2 桁必要です。
- × 0:60　　∵ 分は 0～59 でなくてはなりません。

列 5（製品 ID）の注意点および例

列 5 は、匿名加工データでは、大文字アルファベットと数字の任意の組み合わせによる 1～9 文字の文字列が許されます。元データでは、その一部しか使われていません。

例：

- ○ 0
- ○ M
- × m　　∵ 小文字は使えません。

列 6（単価）の注意点および例

列 6 は正の小数値で、元データは小数点以下 2 桁ですが、匿名加工データは小数点以下の省略が許されます。整数部は 5 桁以下です。

例：

- ○ 99999.99
- ○ 1
- ○ 1.0
- ○ 1.00

- ○ 0.1
- × 1.　　∵ 小数点で終えてはいけません。
- × .1　　∵ 整数部は省略できません。
- × 01234　　∵ 整数部で 0 が先行してはいけません。
- × 0.00　　∵ 正数でなければいけません。

列 7（数量）の注意点および例

列 7 は自然数で、6 桁以下です。

例：

- ○ 999999
- × 012345　　∵ 0 が先行してはいけません。

2016 年 8 月 3 日 Ver. 0.3

PWS CUP 実行委員会