PWS CUP 2018 データの書式

コンテストの加工フェイズで提出する匿名加工データは、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。

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、本書式は少し緩く、たとえば元データの月日は 0 サプレスされていますが本書式では 0 サプレスは任意になっているため、再識別フェイズでは注意してください。

なお、再識別フェイズで受け取る匿名加工データの書式は、次以外は本書式に従っています。

- 行数は、元データの行数以下の任意の整数。
- 行削除を示す行の書式に従う行は含まれない。

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	m 行（m は元データの行数）
各行の文字数	c 文字以下（改行文字は含まない）（c は別途規定）

各行は 5 列（セル）で構成されます。つまり、各行にちょうど 4 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

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

でも

12680,2011/12/9,22138,4.95,3[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）。

行削除（DEL）を示す行の書式

行削除の加工を実施したことを示すには、各列に値「*」を記載してください。次の通りとなります。

,,*,*,*<LF>

各列の書式

各列の書式は次の通りです。

#	通称	書式
1	顧客 ID	9 桁以下の自然数
2	年月日	2010, 2011 年の YYYY/MM/DD 形式 (0 サプレスは任意)、あるいは最小値と最大値をその形式で順に記載した閉区間の形式 [YYYY/MM/DD;YYYY/MM/DD]。あるいはセル削除を示すアスタリスク *。
3	製品 ID	元データの同列に含まれるいずれかの値、あるいはそれらを要素とする集合の形式で {e1;e2;e3} といった形式（空集合や要素の重複は不可、要素数は e 以下（e は別途規定））。あるいはセル削除を示すアスタリスク *。
4	単価	整数部 5 桁以下小数部 2 桁以下の正小数、あるいは最小値と最大値をその形式で順に記載した閉区間の形式 [MIN;MAX]。あるいはセル削除を示すアスタリスク *。
5	数量	6 桁以下の自然数、あるいは最小値と最大値をその形式で順に記載した閉区間の形式 [MIN;MAX]。あるいはセル削除を示すアスタリスク *。

各列とも、値の前後など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（年月日）の注意点および例

列 2 は「年/月/日」あるいはその閉区間、あるいはセル削除を示す「*」です。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。閉区間は大かっこ中に最小値と最大値を順にセミコロン区切りで記載します。

例：

- ○ 2010/1/1
- ○ [2010/1/01;2010/1/31]
- ○ [2010/1/1;2010/01/01]
- ○ *
- × 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。

- × [2010/03/01;2010/2/28] ∵ 最小値、最大値の順にしなくてはいけません。

列 3（製品 ID）の注意点および例

列 3 は元データの同列に含まれるいずれかの値、あるいはその集合、あるいはセル削除を示す「*」です。元データでの値は大文字アルファベットと数字の 1 文字以上の組み合わせで構成されています。集合は中かっこ中に要素をセミコロン区切りで記載します。空集合や要素の重複は不可ですが、要素の順序に制限はありません。

例：

- ○ M （M が元データの同列に含まれている場合）
- ○ {M} （M が元データの同列に含まれている場合）
- ○ {M;22138;C2} （M, 22138, C2 が元データの同列に含まれている場合）
- ○ *
- × m ∵ 元データに小文字は含まれていません。
- × {M;M} ∵ 要素を重複させてはいけません。
- × {M;22138;C2;DOT} （e=3 の場合） ∵ 要素数は e を超えてはいけません。

列 4（単価）の注意点および例

列 4 は正の小数値、あるいはその閉区間、あるいはセル削除を示す「*」です。小数値は、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。閉区間の記法は列 2 と同じです。

例：

- ○ 99999.99
- ○ 1
- ○ [1.00;1.0]
- ○ 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。
- × [10;2.0] ∵ 最小値、最大値の順にしなくてはいけません。

列 5（数量）の注意点および例

列 5 は 6 桁以下の自然数、あるいはその閉区間、あるいはセル削除を示す「*」です。閉区間の記法は列 2 と同じです。

例：

- ○ [1;999999]
- ○ *
- × 012345　　∵ 0 が先行してはいけません。

2018 年 8 月 22 日 Ver. 1

PWS CUP 実行委員会