Transformation de données

Comment transformer un fichier de données non formatté dans en un fichier de données correctement tabulé pour Excel, Stata,... ?

- En utilisant le logiciel **TextPad** (gratuit)
- En utilisant la fonction **remplacement par expressions régulières** de ce logiciel

Démonstration N°1

Par exemple, si vos données sont sous cette forme (et ce, sur plus de 15000 lignes):

```
     2 AK  20      Anchorage Borough        380 MSA: Anchorage, AK
MED AREA MEDIAN FAMILY INCOME (EFFECTIVE 12-14-95):  $ 57100
L50 VERY LOW INCOME (1-8 PER.) 20000 22850 25700 28550 30850 33100 35400 37700
L80 LOW INCOME (1-8 PERSON)    29100 33300 37450 41600 44950 48250 51600 54900
FMR FAIR MARKET RENT(0-4 BEDRM EFFECTIVE 2-22-96)  473   557   740  1029  1215
     2 AK  13      Aleutians East Borough     0 MSA:   (NONMETROPOLITAN AREA)
MED AREA MEDIAN FAMILY INCOME (EFFECTIVE 12-14-95):  $ 52000
L50 VERY LOW INCOME (1-8 PER.) 18200 20800 23400 26000 28100 30150 32250 34300
L80 LOW INCOME (1-8 PERSON)    29100 33300 37450 41600 44950 48250 51600 54900
FMR FAIR MARKET RENT(0-4 BEDRM EFFECTIVE 2-22-96)  491   553   623   778  1020
     2 AK  16      Aleutians West Census      0 MSA:   (NONMETROPOLITAN AREA)
MED AREA MEDIAN FAMILY INCOME (EFFECTIVE 12-14-95):  $ 39000
L50 VERY LOW INCOME (1-8 PER.) 17000 19400 21800 24250 26200 28150 30050 32000
L80 LOW INCOME (1-8 PERSON)    27150 31050 34900 38800 41900 45000 48100 51200
FMR FAIR MARKET RENT(0-4 BEDRM EFFECTIVE 2-22-96)  421   475   533   668   748
```

Hors de question de tout séparer à la main, vous deviendriez fou (ou folle). Nous allons utiliser les expressions régulières pour mettre de l'ordre dans tout ça.

Ouvrez vos données dans TextPad. Je vous conseille de faire des sauvegardes régulières sur des fichiers différents (document1.txt, document2.txt,...) à chaque étapes importantes et réussies de vos transformations.

Cliquez sur F8 et faites les remplacements suivants successivement.

Rechercher:	Remplacer par:	Notes
\n_____	@	(1a)
\n	\t	(1b)
@	\n	(1c)
__	_	refaire autant que possible (2a)
_\([0-9]+\)	\t\1	(2b)
\([0-9]+\)_	\1\t	(2c)
le caractère _ égale le caractère		

d'espacement

IMPORTANT : il faut que la coche "Expression régulière" soit activée.

Explications de texte :

Première étape: mettre chaque enregistrement sur une ligne unique.

Dans notre fichier d'exemple, les données d'un enregistrement se présentent sur 4 lignes. Il faut les ramener les donnée sur une ligne. Pour cela, je repère quelle est la séquence de texte qui sépare deux enregistrements. C'est la séquence "Saut à la ligne suivie de 4 espaces" (1a). Je la remplace par "@" : c'est un caractère qui ne se trouve nulle part dans les données, donc parfait pour séparer les enregistrements.

Ensuite, je remplace tous les autres "Saut de ligne" par des "tabulations" (1b).

Enfin, je rétablis le tout en remplaçant "@" (le caractère spécial que j'ai choisi pour séparer les enregistrements) par un vrai "saut de ligne".

Deuxième étape : Séparer les données par des tabulations

Maintenant que nous avons un enregistrement par ligne, comment allons-nous séparer les différentes données par des tabulations ?

Je remarque que les données sont séparées par un nombre d'espacements très variable. Pour y voir plus clair, je décide de remplacer "tous les espacements multiples" par "un seul" (2a). Pour cela, il faut recommencer cette opération plusieurs fois (jusqu'à ce que le message suivant s'affiche : "impossible de trouver...").

Je pourrais remplacer maintenant tous les espacements par des tabulations mais je remarque qu'il y a des groupes de mots que je ne veux pas séparer. Après une observation plus attentive, je vois qu'il faut que je remplace tout ce qui ressemble de près ou de loin à "un espacement suivi d'un chiffre" par "une tabulation" (2b), et aussi "un chiffre suivi d'un espacement" par "une tabulation" (2c).

Il reste enfin quelques bricoles à normaliser et le fichier est près pour être lu par votre traiteur (de statistiques) préféré.

Pour commentaires ou information : SCECO-information [at] umontreal.ca
Page mise à jour le 22-10-2010