Cas pràctic: manipulació de fitxers de text

PID_00270618

Gerard Farràs Ballabriga

Temps mínim de dedicació recomanat: 1 hora

Universitat Oberta de Catalunya

Gerard Farràs Ballabriga

Enginyer tècnic en Informàtica de sistemes per la Universitat Autònoma de Barcelona (UAB). Enginyer en Informàtica i màster en Societat de la Informació i el Coneixement per la Universitat Oberta de Catalunya (UOC). Actualment treballa com a professor en una escola de secundària i formació professional. Anteriorment ha desenvolupat la seva activitat professional en l'àrea de sistemes d'informació d'un centre tecnològic i també com a professional autònom (*freelance*) treballant com administrador de sistemes i desenvolupador web.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Julià Minguillón Alfonso (2020)

Primera edició: febrer 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

1. Introducció

El Servei Català de la Salut genera un catàleg actualitzat mensualment amb «tots els medicaments dispensables en oficines de farmàcia i els productes sanitaris del Sistema Nacional de Salut que finança el CatSalut mitjançant les receptes mèdiques oficials. Inclou les dades d'identificació del producte, les dades econòmiques i de composició». Aquest fitxer està en un format pla de text i és automatitzable, ja que segueix un format concret.

Catàleg de productes farmacèutics

La informació completa d'aquest catàleg està disponible a l'enllaç: catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/productes-farmaceutics/ (pestanya «Descàrrega del catàleg complet»).

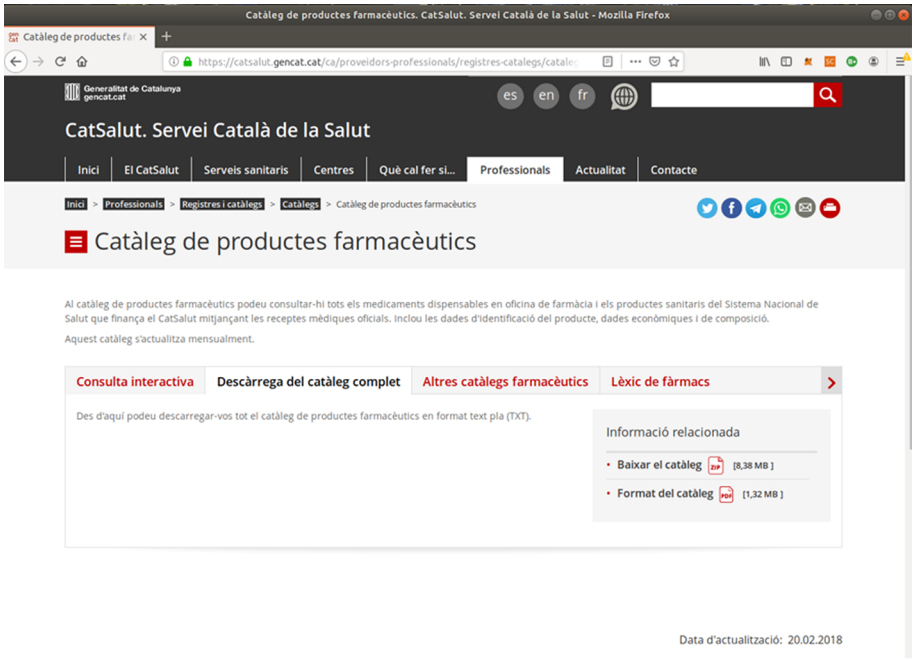
L'objectiu d'aquest exercici consisteix a implementar un sistema automatitzat que, el dia 2 de cada mes, obtingui dos camps concrets de l'apartat nomenclàtors medicaments i productes sanitaris (la «Descripció producte farmacèutic» i el «preu de comercialització») i generi un fitxer csv amb aquests dos valors.

Aquest primer cas ha de servir a l'estudiant com a mostra per a tractar fitxers de text pla, sense formats com ara csv, xml o json. Tampoc no empra cap API externa per a l'obtenció de les dades. Es tracta de la lectura i tractament d'un fitxer gran que conté solament text.

Paraules clau

Fitxers TXT, comandes `wget`, `head`, `cut`, `tail`, `paste`.

Figura 1. Lloc web del Servei Català de la Salut des d'on podeu descarregar el catàleg complet que es treballarà en aquest cas



Font: catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/catalegs/productes-farmaceutics

2. Passos

El primer cas consisteix a descarregar el catàleg i també el format del catàleg per a tenir-lo com a referència.

Nota

Tot i que l'editor de text parteix els enllaços que no caben en una sola línia, es tracta solament d'una comanda.

```
usuari@nomMaquina:~$ wget -q https://catsalut.gencat.cat/web/.content/minisite/catsalut/
proveidors_professionals/registres_catalegs/documents/catalegfarmacia.zip

usuari@nomMaquina:~$ ls -lh catalegfarmacia.zip
-rw-r--r-- 1 usuari usuari 8,4M de se 3 13:04 catalegfarmacia.zip
```

Observem que aquest fitxer comprimit ocupa 8,4 megabytes. El podem descomprimir amb la comanda següent:

```
usuari@nomMaquina:~$ unzip catalegfarmacia.zip
Archive: catalegfarmacia.zip
inflating: CATALEGFARMACIA20190901.TXT
```

També serà útil disposar del format del catàleg, per a tenir-lo com a referència:

```
usuari@nomMaquina:~$ wget -q https://catsalut.gencat.cat/web/.content/minisite/catsalut/
proveidors_professionals/registres_catalegs/documents/for_extrac_cpf.pdf
```

El fitxer descomprimit és un text pla sense elements separadors de cada camp, ja que cadascun d'aquests s'especifica en una posició concreta (es recomana fer un cop d'ull a les primeres pàgines del format del catàleg). Podem obtenir informació d'aquest fitxer amb les comandes següents:

```
usuari@nomMaquina:~$ file CATALEGFARMACIA20190901.TXT
CATALEGFARMACIA20190901.TXT: ISO-8859 text, with very long lines, with CRLF line terminators

usuari@nomMaquina:~$ wc -l CATALEGFARMACIA20190901.TXT
511540 CATALEGFARMACIA20190901.TXT
```

Aquest fitxer concret té més de mig milió de línies. Tal com especifica el format del catàleg, el primer registre conté la capçalera del fitxer:

```
usuari@nomMaquina:~$ head -1 CATALEGFARMACIA20190701.TXT
001 PFC00013S20190701 2019062810091200473879000043 2019HPIT3PFC PFC 00013 0001300130013000000
```

I, en el segon, el registre de capçalera de detall dels nomenclàtors dels medicaments i productes sanitaris:

```
usuari@nomMaquina:~$ head -2 CATALEGFARMACIA20190701.TXT | tail -1
100000101 2019070120190630201906281009120006311 5PFC18001SNOMENCLATOR
EF NORMALS I EFECTES / ACCESORIS
```

El nombre de registres del grup que depenen de la capçalera de detall està en aquesta línia (es tracta d'una xifra de vuit caràcters. El camp concret s'anomena «Nombre de registres del grup que depenen de la capçalera de detall»). Obtenim aquest valor amb la comanda següent:

```
usuari@nomMaquina:~$ d=`head -2 CATALEGFARMACIA20190901.TXT | tail -1 | cut -c 55-62`
```

Aquesta comanda executa el *head*, que mostra les dues primeres línies del fitxer, les traspassa a un *pipe*, on el *tail* es quedarà amb la darrera línia i, finalment, el passa amb un altre *pipe* a la comanda `cut` que retallarà els caràcters del 55 al 62. Tot plegat s'emmagatzemarà en una variable que hem anomenat `d`. Amb la comanda següent podem observar el valor que hem obtingut:

```
usuari@nomMaquina:~$ echo $d
00063353
```

Cal tractar `$d` línies encara que recordem que les dues anteriors contenien registres de capçalera. Per tant:

```
n=`expr $d + 2`
```

Filtrem ara pel camp «Descripció producte farmacèutic» (que està, segons el catàleg, en els valors 22 i 121 i ocupa un total de 100 caràcters):

```
usuari@nomMaquina:~$ head -$n CATALEGFARMACIA20190901.TXT | tail -$d | cut -c22-121
```

Hem obtingut un dels camps que estàvem buscant.

Aquest resultat conté més de 60.000 línies (en el moment de fer aquest exercici, concretament, 63.115, encara que és una xifra que pot variar). A tall de demostració, mostrarem les 15 primeres:

```
usuari@nomMaquina:~$ head -$n CATALEGFARMACIA20190901.TXT | tail -$d | cut -c22-121 | head -15
APALOZ 5 MG COMPRIMIDOS EFG , 28 comprimidos
MEDIA LARGA (A-F) COMP NORMAL KURVAY-B (500 TALLA PEQUEÑA
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH12 B10
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH14 B10
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH16 B10
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH18 B10
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH20 B10
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH22 B10
SONDA VESICAL SILICONA FOLEY SILICONA 100% CH24 B10
MEDIA LARGA (A-F) COMP NORMAL KURVAY-B (500 TALLA MEDIANA
BOLSAS ILEOST RES SINT MIC FIL MODERMA FLEX ABIERTA PLANA MINI OPACA 15-55MM 30U
BOLSAS ILEOST RES SINT MIC FIL MODERMA FLEX ABIERTA CONVEXA MAXI OPACA 15-38MM 30U
BOLSAS ILEOST RES SINT MIC FIL MODERMA FLEX ABIERTA CONVEXA MAXI OPACA 15-51MM 30U
BOLSAS ILEOST RES SINT MIC FIL MODERMA FLEX ABIERTA CONVEXA MAXI TRANSPARENTE 15-38MM 30U
BOLSAS ILEOST RES SINT MIC FIL MODERMA FLEX ABIERTA CONVEXA MAXI TRANSPARENTE 15-51MM 30U
```

De manera similar obtindrem un altre camp: «Preu de comercialització», que s'expressa amb vuit números (sis valors enters i dos decimals) i està entre la posició 131 i 138:

```
usuari@nomMaquina:~$ head -$n CATALEGFARMACIA20190901.TXT | tail -$d | cut -c131-138 | head -15
00000000
00000652
00000809
00000809
```

```
00000809
00000809
00000809
00000809
00000809
00000652
00006850
00006850
00006850
00006850
00006850
```

A fi de generar un fitxer .csv amb la informació dels dos camps, els mesclarem amb la comanda `paste`. Per a fer-ho, abans generarem un parell de fitxers auxiliars amb la informació de cada camp:

```
usuari@nomMaquina:~$ head -$n CATALEGFARMACIA20190901.TXT | tail -$d | cut -c22-121 > descripcio.txt
usuari@nomMaquina:~$ head -$n CATALEGFARMACIA20190901.TXT | tail -$d | cut -c131-138 > cost.txt
```

Amb la comanda següent és possible mesclar el contingut dels dos fitxers (aquí solament es mostra la primera línia, a tall d'exemple):

```
usuari@nomMaquina:~$ paste -d ';' descripcio.txt cost.txt | head -1
APALOZ 5 MG COMPRIMIDOS EFG , 28 comprimidos ;00000000
```

El paràmetre `-d` indica el separador que desitgem.

Si es volgués protegir amb cometes ("), en els dos camps es podrien fer substitucions similars a les següents:

```
usuari@nomMaquina:~$ paste -d ';' descripcio.txt cost.txt |
sed s/^/\"/g | sed s/$/\"/g | sed s/\;/\"\;\"/g > dades.csv
```

En el primer `sed` se substitueix l'inici de cada línia agregant unes cometes (recordem que s'indica l'inici de cada línia fent servir el caràcter especial `^`), en el segon el mateix però al final de línia (que s'indica amb el caràcter `$`) i, en el darrer, se substitueix el caràcter separador que havíem agregat `;` , per `;`. Recordem que les `\` serveixen per a protegir caràcters especials.

Quedaria solament automatitzar aquest procés per tal que solament s'executés el dia 2 de cada mes a una hora específica.

Primer caldria agregar tots els passos en un sol *script*. Aquest guió suprimeix primer els possibles fitxers resultants d'execucions anteriors. El nom del fitxer també varia cada mes (malgrat comença sempre per «CATALEGFARMACIA» i finalitza amb un «.TXT» es canvien les xifres que indiquen la data). Per tant, caldrà obtenir el nom del fitxer en concret. Recordem que les línies que comencen per un coixinet # actuen com a comentaris i es poden obviar si s'escriu l'*script*.

Hem anomenat aquest *script* «obteDades.sh» i es mostra el codi a continuació:

```
#!/bin/bash
# Script per a tractar dades.
```

```
# Accedim primer a la carpeta on desitgem realitzar tot el tractament.
cd /home/usuari/

# Suprimim possibles fitxers anteriors.
# El paràmetre -f força a no demanar res a l'usuari.
rm -f catalegfarmacia.zip
rm -f CATALEGFARMACIA*.TXT
rm -f descripcio.txt
rm -f cost.txt
rm -f dades.csv

# Obtenim el fitxer i el descomprimim.
wget https://catsalut.gencat.cat/web/.content/minisite/catsalut/proveidors_professionals/
registres_catalegs/documents/catalegfarmacia.zip

unzip catalegfarmacia.zip

# Obtenim el nom del fitxer concret, ja que varia cada mes.
# S'ubica en una variable d'entorn.
nomfitxer=`ls CATALEGFARMACIA*.TXT`


# Obtenim el nombre de registres a tractar.
d=`head -2 $nomfitxer | tail -1 | cut -c 55-62`
n=`expr $d + 2`

# Generem un fitxer diferent per a cada camp.
head -$n $nomfitxer | tail -$d | cut -c22-121 > descripcio.txt
head -$n $nomfitxer | tail -$d | cut -c131-138 > cost.txt

# Unim ambdós fitxers i generem un fitxer .csv
paste -d ';' descripcio.txt cost.txt | sed s/^/\"/g | sed s/$/\"/g | sed s/\;/\"\;\"/g > dades.csv
```

Recordem agregar permisos d'execució en aquest fitxer amb la comanda `chmod`:

```
usuari@nomMaquina:~$ chmod +x obteDades.sh

usuari@nomMaquina:~$ ls -l obteDades.sh
-rwxr-xr-x 1 usuari usuari 1101 de se 27 06:32 obteDades.sh
```

Finalment, caldria agregar una línia en el fitxer «/etc/crontab» que automatitzaria l'execució:

```
30 7    2 * *   usuari    /home/usuari/obteDades.sh
```

A les 7:30 del dia 2 de cada mes, en aquesta carpeta hi haurà un fitxer «dades.csv» amb els dos camps del catàleg. Per exemple:

```
usuari@nomMaquina:~$ head -10 dades.csv
"APALOZ 5 MG COMPRIMIDOS EFG , 28 comprimidos ";"00000000"
"MEDIA LARGA (A-F) COMP NORMAL KURVAY-B (500 TALLA PEQUE#A ";"00000652"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH12 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH14 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH16 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH18 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH20 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH22 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH24 B10 ";"00000809"
"MEDIA LARGA (A-F) COMP NORMAL KURVAY-B (500 TALLA MEDIANA ";"00000652"
```

Observem que és possible que hi hagi algun problema amb la codificació de caràcters. Concretament, no es llegeix adequadament el caràcter «Ñ» (segona línia, on mostra «TALLA PEQUEÑA»). Caldria canviar la codificació dels caràcters emprant la comanda següent:

```
usuari@nomMaquina:~$ iconv -t UTF-8 -f ISO-8859-1 dades.csv | head -10
"APALOZ 5 MG COMPRIMIDOS EFG , 28 comprimidos ";"00000000"
"MEDIA LARGA (A-F) COMP NORMAL KURVAY-B (500 TALLA PEQUEÑA ";"00000652"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH12 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH14 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH16 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH18 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH20 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH22 B10 ";"00000809"
"SONDA VESICAL SILICONA FOLEY SILICONA 100% CH24 B10 ";"00000809"
"MEDIA LARGA (A-F) COMP NORMAL KURVAY-B (500 TALLA MEDIANA ";"00000652"
```