PEC 3: Expresiones regulares con grep y sed

El objetivo de esta PEC consiste en la utilización de expresiones regulares, mediante el comando `grep` para buscar regularidades en varios tipos de documentos y el comando `sed` para su modificación automática. También se usará el comando `awk` para filtrar datos y realizar pequeños informes.

Para la elaboración de esta PEC es necesario utilizar el usuario creado en la PEC1 (aunque es posible utilizar cualquier otra máquina Ubuntu 20.04). Deben demostrarse las operaciones a través de capturas de pantalla con el usuario creado en la PEC1, de otro modo se considerará una PEC plagiada.

*Formato de entrega: fichero en formato zip que contenga un fichero de texto llamado PEC3_2.txt, cada uno de los scripts pedidos y las capturas de pantalla de su ejecución (**sin recortar,** donde **se vea claramente el usuario de la UOC creado en la PEC1**, y un **máximo de 3 capturas por ejercicio**) con la nomenclatura siguiente:*

- *Capturas de pantalla: en los ejercicios donde se indique explícitamente, se usará la nomenclatura **PEC3_NumeroEjercicio_Apartado_NumeroCaptura.png**. Si el ejercicio no contiene apartados, entonces el apartado debe ser omitido. Por ejemplo: PEC3_1_1.png, PEC3_1_2.png, PEC3_2_a_1.png, PEC3_4_a_1.png, PEC3_4_a_2.png.*
- *Fichero de texto PEC3_2.txt*
- Los scripts pedidos:
  - PEC3_1.sh
  - PEC3_3_a.sh
  - PEC3_3_b.sh
  - PEC3_4_a.awk
  - PEC3_4_b.awk

NOTA importante: en una asignatura donde la automatización de tareas es el objetivo, respetar todos los nombres de los scripts es muy importante, vigilad el uso de mayúsculas/minúsculas, guiones, etc.

Enunciado

Para el desarrollo de esta PEC3 vamos a utilizar en exclusiva el fichero adjunto con el enunciado denominado **demographic_info.csv** del conjunto de datos "Improving Artificial Teachers by Considering How People Learn and Forget: Dataset" (https://zenodo.org/record/5536917). El dataset es descargable a través del enlace https://zenodo.org/record/5536917/files/demographic_info.csv?download=1, pero para el desarrollo de la PEC3 deberá hacerse en base al fichero que se adjunta.

El nombre del fichero de datos no debe cambiarse. Todos los archivos deben estar en la misma carpeta.

Ejercicio 1

Puntuación: 2 puntos

A partir del fichero adjunto al enunciado de esta PEC3, llamado **demographic_info.csv**, tenéis que encontrar **una única expresión regular** del tipo **ERE** que devuelva los registros que cumplan con las siguientes condiciones:

- El campo con el identificador (user) sea mayor o igual a 10 y menor que 99.
- Sea relativo exclusivamente a mujeres (sexo femenino).
- Edad comprendida entre los rangos de edad de 10 a 19 años y de 30 a 39, ambos inclusive.
- La lengua nativa sea el finlandés y además sepan también hablar el español y el francés.

La expresión regular se debe usar dentro de un script en Bash llamado *PEC3_1.sh* que ejecute un comando **`grep`**, teniendo en cuenta como se ha especificado, y que la expresión sea del tipo ERE.

En la entrega debe adjuntarse el script y al menos una captura con su ejecución.

Ejercicio 2

Puntuación: 1.5 puntos

Dominar las expresiones regulares incluye también la capacidad de comprender y modificar expresiones regulares **creadas por otros programadores**. De hecho, es lo más habitual. Este ejercicio está centrado en la comprensión y modificación, si es el caso, de **expresiones regulares ya dadas**. A partir de la siguiente expresión regular:

```
^[0-9]*,(M?|F?),[0-9]*,[a-zA-Z]*,$
```

Se pide:

- Explica la utilidad (sin describir técnicamente la estructura de la expresión regular), que se pretende obtener con la expresión regular anterior:
- ¿Qué tipo de expresión regular es? Justifica muy brevemente la respuesta.

Escribid vuestras respuestas en un fichero de texto plano llamado *PEC3_2.txt*

Ejercicio 3

Puntuación: 3 puntos

Este ejercicio sirve para **ilustrar y mostrar el uso del editor `sed`**. Éste es especialmente útil para hacer sustituciones de expresiones regulares, aunque sus funcionalidades no se limitan a éso. En este ejercicio tendréis que poner en práctica el uso de `sed` para diferentes tareas.

A partir del dataset adjunto a la PEC3, es decir, del fichero **demographic_info.csv**, lleva a cabo las siguientes acciones:

- Crea un script en bash denominado *PEC3_3_a.sh* que ejecute un **único comando `sed`** que permita **sustituir la letra del campo**

Gender por el nombre completo del sexo (F = Female; M = Male). Si el valor de la columna de entrada es distinto a F ó M entonces **la línea entera deberá omitirse** (no debe mostrarse). *(1.5 puntos).* El script debe funcionar usando la siguiente invocación:

```
./PEC3_3_a.sh
```

- A partir del dataset adjunto se debe crear un script en bash llamado *PEC3_3_b.sh* el cual, mediante el uso de **un único comando `sed`**, sirva para cambiar **de minúsculas a mayúsculas** todos los valores de la columna **native_lang**. *(1.5 puntos).* El script debe funcionar usando la siguiente invocación:

```
./PEC3_3_b.sh
```

En la entrega debe adjuntarse cada script y al menos una captura por script de su ejecución.

Ejercicio 4

Puntuación: 3.5 puntos

Como se ha visto en los materiales docentes, `awk` es una herramienta con muchas posibilidades para filtrar datos, generar informes, etc. Es, de hecho, un lenguaje de programación, por lo que se pueden crear scripts `awk` para automatizar tareas. En este ejercicio se tendrán que crear dos scripts en `awk` a partir del fichero adjunto a la PEC3, **demographic_info.csv**:

a) Realiza un script en `awk` llamado *PEC3_4_a.awk* que calcule **el porcentaje** de personas cuya **lengua nativa es el finlandés y que hablen sueco**. El script tendrá que devolver un número real con una precisión de **dos decimales** seguido del sufijo %. Por ejemplo: 66.55% *(1.75 puntos).* El script debe funcionar usando la invocación:

```
gawk -f PEC3_4_a.awk demographic_info.csv
```

b) Realiza un script en `awk` llamado *PEC3_4_b.awk* que calcule **la edad media** de todos los registros que los que hablan **español**, ya sea lengua nativa o no lo sea. El script tendrá que devolver un número real

con una precisión de **dos decimales**. Por ejemplo: 15.17 *(1.75 puntos).* El script debe funcionar usando la siguiente invocación:

```
gawk -f PEC3_4_b.awk demographic_info.csv
```

En la entrega debe adjuntarse cada script y al menos una captura por script de su ejecución.

Formato de entrega

Entregar un único fichero en formato zip que contenga:

- Capturas de pantalla de la ejecución de cada uno de los scripts pedidos:
  - Mejor si están sin recortar y se lee el resultado (si es posible, completo) de la ejecución del comando.
  - Se incluye como máximo un total de 3 capturas de pantalla por ejercicio.
  - Debe aparecer siempre claramente el usuario de la UOC creado en la PEC1. Las capturas de pantalla donde no aparezca el nombre de usuario creado en la PEC anterior invalidarán el ejercicio completamente (no recibirán puntuación alguna).
  - *Es necesario respetar la nomenclatura de los scripts y de las capturas indicada con anterioridad.*
- Cada uno de los ficheros correspondientes a cada script pedido.
- El fichero *PEC3_2.txt*

Para los ejercicios 1,3 y 4 no se puntuará ningún apartado en donde falte la captura o el código. Ambos son **IMPRESCINDIBLES**.

NOTA: No se aceptará en ningún caso no justificado la entrega de la PEC3 después de la fecha máxima de entrega (*28-11-2021 a las 23:59:59). Si por alguna razón pensáis que no vais a poder entregar a tiempo, consultadlo con vuestro profesor siempre con anterioridad.*