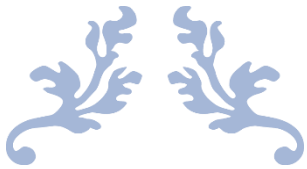


PRÁCTICA 1: SHELL SCRIPT

Julián Blanco González

27 DE OCTUBRE DE 2023

GRADO EN INGENIERÍA INFORMÁTICA

Grupo Lunes / Convocatoria: octubre

Índice

1. Memoria Explicativa de la práctica

1.1. Introducción de la Práctica

La práctica consiste en aplicar un algoritmo para comprobar que si un correo tiene potencial de ser spam o no. Hay un archivo donde están los correos a analizar (*Emails.txt*) y un archivo donde están las palabras de spam (*sword.txt*).

Tiene un menú donde hay 5 opciones:

1- Análisis: ver las coincidencias de cada palabra en cada correo y guardarlo en un fichero resultado (.freq).

2- Predicción: aplicando los diferentes cálculos, predice si un correo es potencial de spam o ham.

3- Informes: visualizar diferentes consultas por pantalla: número de términos en cada correo electrónico, ver para un término individual, cada correo en el que aparece (50 primeros caracteres) y ver para un correo individual, todos los términos que aparecen en él.

4- Ayuda: información que se le da al usuario para que pueda entender y utilizar la práctica.

5- Salir: sale de la aplicación.

1.2. Solución aportada

Para almacenar las coincidencias en el apartado 1, se genera una matriz de N x M:

N (filas): número de correos

M (columnas): idCorreo + columna SPAM/HAM (campo 3 del archivo de correos) + número de términos + número total de palabras por correo

IDCORREO	SPAM/HAM	T1	T2	TN	TOTALWORDS
1	1	2	0	3	20
2	0	3	0	5	30
3	1	4	2	6	15

Para poder hacer ahora la predicción, hay que seguir una serie de pasos:

1- Recuperar la matriz desde el archivo de frecuencias

2- Calcular el TF de cada término en cada correo:

TF = (Nº veces término en un correo/ Total Palabras de un correo)

3- Calcular el IDF de cada término en el total de correos:

Para poder hacerlo, primero hay que recorrer por columnas para ver todas las celdas donde un término no es 0. Cuando no sea 0, se suma 1 a la fila nueva creada abajo del todo. Luego, se aplica la siguiente fórmula a cada término:

IDF = log10 (Nº total de correos/Nº de correos que contienen el término)

4- Calcular TF x IDF: teniendo la fila de abajo del todo (IDF) y teniendo cada celda con su correspondiente TF, hacer la multiplicación.

5- Hacer la media de TF x IDF: calcular la media de cada Tf x IDF de cada correo.

6- Comparar el resultado con un valor para ver si es SPAM o HAM: si el resultado de la media es mayor de 0.3, será Spam, si no, es HAM. Se añade el resultado en una columna adicional, así se puede comparar con la predicción inicial.

1.3. Retos que han surgido

Conseguir hacer una condición para ver si el valor de una posición de la matriz es diferente de 0. El problema ha surgido cuando a la hora de hacer el if, como el valor puede ser float, hay que usar el comando bc, y me costó encontrar la manera adecuada. Lo conseguí de esta manera:

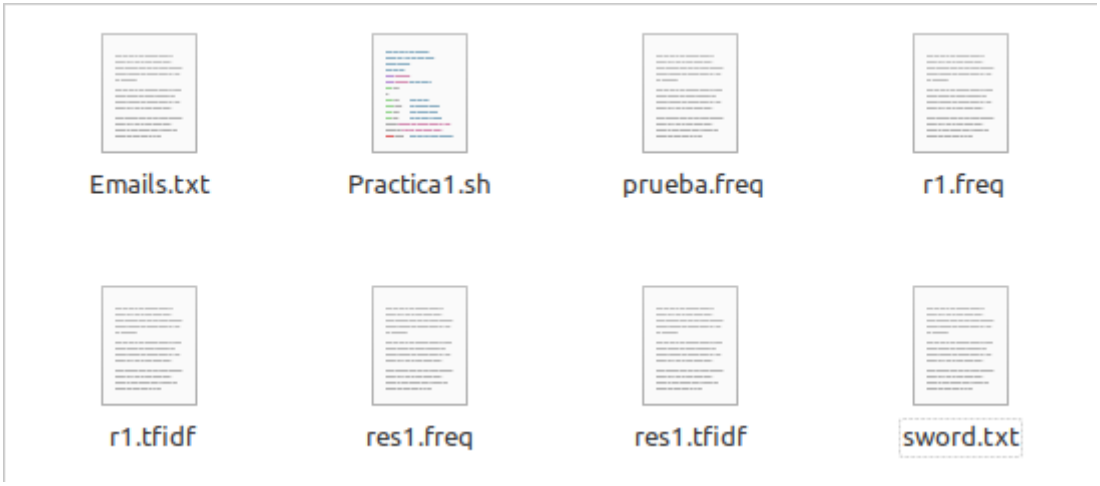
```
valor_matriz="${matriz2[$i,$j]}"

if [ "$(echo "$valor_matriz != 0.0" | bc -l)" -eq 1 ]; then
```

Conseguir controlar todos los errores posibles, ya sea, que el usuario introduzca siempre los datos correctos, y si no lo hace, se muestra un mensaje, a la hora de hacer la división, controlar que nunca se hará por 0…

2. Manual de Usuario

2.1. Carpeta del programa

Al descargar la carpeta, se abrirá unos archivos como los siguientes:



- .tfidf: archivo de predicción realizada (se generan al usar opción 2)
- .freq: archivo de frecuencias (se generan al usar opción 1)
- Emails.txt: archivo de correos.
- sword.txt: archivo de palabras.
- Practica1.sh: ejecutable shell de la práctica.

2.2. Como ejecutar la aplicación

Para ejecutar la aplicación, hay que seguir los siguientes pasos:

1- Abrir la consola de Linux

```
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
julian@julian-VirtualBox:~$
```

2- Con el comando cd, ir a la ruta donde se haya descargado la carpeta

```
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
julian@julian-VirtualBox:~$ cd Escritorio/1_SO/
julian@julian-VirtualBox:~/Escritorio/1_SO$
```

3- Para ejecutar la práctica, hay 2 maneras: usar bash + NombreEjecutable.sh o usar ./NombrePráctica.sh. Para este último, habría que darle permisos de ejecución y escritura: chmod u+x NombrePráctica.sh

```
julian@julian-VirtualBox:~$ cd Escritorio/1_SO/
julian@julian-VirtualBox:~/Escritorio/1_SO$ ./Practica1.sh
```

2.3. Manual de la Aplicación

Página principal de la aplicación: el usuario puede elegir una de las 5 opciones:

```
Archivo   Editar   Ver   Buscar   Terminal   Ayuda
----------Menú de la Aplicación----------
1. Análisis de datos
2. Predicción
3. Informes de resultados
4. Ayuda
5. Salir de la aplicación
Opción: 
```

Opción 1: introducir el fichero de datos, introducir el fichero de corros e introducir el fichero de resultados (.freq):

```
1. Análisis de datos
=========================================================================================================
En esta opción se pide 3 nombres:
        * Nombre del fichero en el que se encuentran las palabras o términos a buscar: (sword.txt)
        * Nombre del fichero donde se almacenan los correos electrónicos: (Emails.txt)
        * Nombre del fichero donde se desea almacenar el resultado del análisis
=========================================================================================================
1 --> Nombre del fichero de palabras
Por favor, introduce el nombre del fichero de palabras: sword.txt
El fichero sword.txt existe.
====================================================
2 --> Nombre del fichero de correos
Por favor, introduce el nombre del fichero de correos: Emails.txt
El fichero  existe.
====================================================
3 --> Nombre del fichero para guardar los resultados
Por último, introduce el nombre del fichero resultados(solo el nombre): r1
Se ha creado el archivo r1.freq en la carpeta actual.
numero de filas: 4
numero de columnas: 8
Total palabras correo 1: 10
Total palabras correo 2: 5
Total palabras correo 3: 5
Total palabras correo 4: 17
Pulsa para continuar...
```

Opción 2: predicción

1: hacer la predicción del archivo generado de la opción 1.

```
2. Predicción
==============================================================================================
Opción que calcula si un correo tiene spam o no. Se pide cargar un fichero de frecuencias ya creado o justo el generado por la opción 1 anteriormente:
        1 --> Opción para cargar la matriz después de hacer la opción 1
        2 --> Opción para cargar la matriz desde un archivo de frecuencias (.freq)
==============================================================================================
Introduce 1, 2 o 3(salir): 1
Has elegido 1
Hay que ver si se ha usado la opcion 1 previamente
Se ha utilzado la opcion 1: r1.freq
numero de filas: 4
numero de columnas: 9
Número de filas en la función: 4
Número de columnas en la función: 9
Nombre del fichero de frecuencias: r1
Visualizar la matriz cargada
  1|   |  3|  2|  0|  0|  0|  0| 10|
  2|   |  0|  0|  0|  0|  0|  0|  5|
  3|   |  1|  0|  2|  1|  1|  0|  6|
  4|   |  0|  0|  3|  1|  0|  0| 17|
Total palabras correo 1: 10
Total palabras correo 2: 5
Total palabras correo 3: 6
Total palabras correo 4: 17
TFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
  1|   |.30|.20|  0|  0|  0|  0|
  2|   |  0|  0|  0|  0|  0|  0|
  3|   |.16|  0|.33|.16|.16|  0|
  4|   |  0|  0|.17|.05|  0|  0|
Calcular el IDF de cada término. Para ello, calcular para cada columna, todas las celdas que no sean 0.
```

2: hacer la predicción de un archivo de frecuencias.

```
2. Predicción
================================================================================================
Opción que calcula si un correo tiene spam o no. Se pide cargar un fichero de frecuencias ya creado o justo el generado por la opción 1 anteriormente:
        1 --> Opción para cargar la matriz después de hacer la opción 1
        2 --> Opción para cargar la matriz desde un archivo de frecuencias (.freq)
================================================================================================
Introduce 1, 2 o 3(salir): 2
Has elegido 2
Cargar el fichero de frecuencias con extensión .freq
Introduce el nombre del archivo de frecuencias (sin la extensión .freq): r1
Existe
No existe el archivo predicción
Número de filas en la función: 4
Número de columnas en la función: 9
Nombre del fichero de frecuencias: r1
Visualizar la matriz cargada
```

Opción 3: diferentes informes que se pueden realizar

Informe 1: número de coincidencias de cada término en todos los correos electrónicos.

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 1
Opción 1 escogida
1- Informe de formato fila/columna de un término.
Introduce el nombre del fichero de frecuencias(con la extensión): r1.freq
Existe el archivo r1.freq
numero de filas: 4
numero de columnas: 9
  1|  0|  2|  2|  1|  0|
Término                                          Nºveces que aparece
------------------------------------------------------------------------------------
Hola                                              1
Adiós                                            0
Malo                                              2
Spam                                              2
Horroroso                                         1
asdsad                                            0
Pulsa para continuar...
```

Informe 2: introduce un término, y muestra todos los correos donde existe esa palabra (solo los 50 primeros caracteres).

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 2
Opción 2 escogida
2- Informe de un término en particular
Introduce el nombre del fichero de frecuencias(con la extensión): r1.freq
Introduce el término a buscar en minúsculas y sin caracteres especiales: hola
Buscando hola en el archivo Fraud_words.txt
numero de filas: 4
numero de columnas: 9
Visualizar la matriz cargada
  1|   |  3|  0|  0|  0|  0|  0|
  2|   |  0|  0|  0|  0|  0|  0|
  3|   |  0|  0|  2|  1|  1|  0|
  4|   |  0|  0|  3|  1|  0|  0|
hola encontrado en la posición: 0 de Fraud_words.txt
Verificando fila 1, posición 1: 3
este es: 3
entra
Verificando fila 2, posición 1: 0
este es: 0
Verificando fila 3, posición 1: 0
este es: 0
Verificando fila 4, posición 1: 0
este es: 0
Posiciones: 1 | Buscando en el archivo Emails.txt si aparece el término...
Está en 1 correos
Índice del correo   Contenido del correo
------------------------------------------------------------------------
1                    1|Hola para ver si funciona esta Hola Hola Adios 1
Pulsa para continuar...
```

Informe 3: introduce un id, y muestra para ese correo, los términos encontrados en ese correo.

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 3
Opción 3 escogida
3- Informe número de términos en un correo
Introduce el nombre del fichero de frecuencias(con la extensión): r1.freq
r1.freq existe.
Introduce el identificador del correo: 3
Buscando 3 en el archivo en la matriz
numero de filas: 4
numero de columnas: 9
Visualizar la matriz cargada
  1|   |  3|  0|  0|  0|  0|  0|
  2|   |  0|  0|  0|  0|  0|  0|
  3|   |  0|  0|  2|  1|  1|  0|
  4|   |  0|  0|  3|  1|  0|  0|
El identificador pertenece a un correo.
Posiciones: 3 | 4 | 5 |
Hay un total de 3 palabras de spam en el correo con id: 3
Los términos son: Malo  |  Spam  |  Horroroso  |
Pulsa para continuar...
```

Opción 4: Manual de ayuda

3. Manual del Programador

3.1. Bucle para pedir solo números enteros al usuario

```
while true; do
    echo -n "Introduce un número (1, 2, 3 o 4 (salir)): "
    read numero

    if [ -z "$numero" ]; then
        echo "No has ingresado nada. Por favor, introduce un número."
    elif [[ ! "$numero" =~ ^[0-9]+$ ]]; then
        echo "Entrada no válida. Por favor, introduce un número válido."
    # Verificar si el número es 1, 2 o 3
    elif [ "$numero" -eq 1 ] || [ "$numero" -eq 2 ] || [ "$numero" -eq 3 ] || [ "$numero" -eq 4 ]; then
        break  # Salir del bucle
    else
        echo "Número no válido. Por favor, introduce 1, 2, 3 o 4."
    fi
done
```

Compruebo primero si se teclea algo, segundo si es un número, y tercero, que sea un número de las opciones posibles, si no, pide un número nuevo.

3.2 Contar las filas y columnas de una matriz en un archivo

```
nfilas=`cat $ficheroFrec_2 | wc -l`
echo "numero de filas: $nfilas"
ncolumnas=`head -n 1 $ficheroFrec_2 | awk -F ":" {'print NF'}`
#ncolumnas=$((ncolumnas - 3))
echo "numero de columnas: $ncolumnas"
```

Con el comando cat, se abre el fichero en cuestión y con el comando wc -l, se cuenta el número de líneas del archivo. Se unen mediante el pipe.

Para las columnas, se utiliza el comando head para coger la primera línea del archivo, y luego, con el comando awk, se coge cada campo del archivo que esté separado por el delimitador ":".

3.3 Fila adicional para IDF

```
for ((j=3;j<=ncolumnas;j++)) do #Inicializar a 0 lal fila nueva
    matriz2[$((nfilas+1)),$j]=0
done

n=1
for ((i=1;i<=nfilas;i++)) do
    for ((j=3;j<=ncolumnas;j++)) do
        valor_matriz="${matriz2[$i,$j]}"
        if [ "$(echo "$valor_matriz != 0.0" | bc -l)" -eq 1 ]; then  #Ver las celdas que son diferente de 0
            echo "valor matriz = $valor_matriz"
            matriz2[$((nfilas+1)),$j]=$((${matriz2[$((nfilas+1)),$j]} + 1)) #sumo 1 a la fila que creo para el IDF
            echo "Entra $n veces"
            n=$((n+1))
        fi

    done
done
```

Primer, inicializo esa fila a 0.

Luego recorro la matriz entera, y siempre que haya una celda que no sea 0, sumo 1 a la posición de la columna en cuestión en la fila nueva.

3.4 Función de cálculo de TF-IDF

```
function calcular_tfidf(){


    nfilas=$1  #Recuperar los 4 parámetros
    ncolumnas=$2
    matriz2=$3
    nombreFicheroFrecuencia=$4


    echo "Número de filas en la función: $nfilas"
    echo "Número de columnas en la función: $ncolumnas"
    echo "Nombre del fichero de frecuencias: $nombreFicheroFrecuencia"


    echo "Visualizar la matriz cargada"
    for ((i=1; i<=nfilas; i++)); do
        for ((j=1; j<=ncolumnas; j++)); do
            printf "%3s|" "${matriz2[$i,$j]}"
        done
        echo
    done

    #-calcular el TF de cada término en cada correo
    ncolumnas=$((ncolumnas - 1))
    for ((i=1;i<=nfilas;i++)) do
         total_palabras="${matriz2[$i,$((ncolumnas+1))]}"
         echo "Total palabras correo $i: $total_palabras"
         for ((j=3;j<=ncolumnas;j++)) do
              tf=$(echo "scale=2; ${matriz2[$i,$j]} / $total_palabras" | bc)
              matriz2[$i,$j]=$tf
         done
    done
```

La función tiene 4 parámetros:

- El número de filas: filas de la matriz
- El número de columnas: columnas de la matriz

- La matriz: hacer los cálculos
- El archivo de frecuencias: guardar el fichero con extensión .tfidf

4. Juego de Pruebas

Voy a utilizar los siguientes archivos de correo y palabras para hacer las pruebas. También voy a utilizar "echos" para ver el progreso al hacer cada opción del menú.

4.1. Datos entrada

Archivo de correos (Emails.txt): tiene 8 correos.

```
1|Oferta especial solo por hoy! Compra ahora y ahorra dinero|1|
2|Reunión de equipo a las 3 PM en la sala de conferencias|0|
3|Gana un viaje gratis a destinos exóticos. ¡Regístrate ahora!|1|
4|Recordatorio: Pago de factura pendiente de $100|0|
5|Descuento del 20% en tu próxima compra con nosotros|1|
6|¡Felicidades! Eres el ganador de nuestro concurso mensual|1|
7|Actualización de política de privacidad y términos de servicio|0|
8|Descuento del 20% en tu próxima compra con nosotros|1|
```

Archivo de palabras (sword.txt): tiene 6 palabras.

```
Oferta
Gratis
Gana
Descuento
Felicidades
Registro
```

4.2. Opción 1

Ejecución buena: introduciendo los 3 nombres y comprobando todo

```
1. Análisis de datos
=====================================================================================================
En esta opción se pide 3 nombres:
        * Nombre del fichero en el que se encuentran las palabras o términos a buscar: (sword.txt)
        * Nombre del fichero donde se almacenan los correos electrónicos: (Emails.txt)
        * Nombre del fichero donde se desea almacenar el resultado del análisis (sin extensión)
=====================================================================================================
1 --> Nombre del fichero de palabras
Por favor, introduce el nombre del fichero de palabras: sword.txt
=====================================================
2 --> Nombre del fichero de correos
Por favor, introduce el nombre del fichero de correos: Emails.txt
=====================================================
3 --> Nombre del fichero para guardar los resultados
Por último, introduce el nombre del fichero resultados(solo el nombre): resultado
Se ha creado el archivo resultado.freq en la carpeta actual.
=====================================================
Análisis realizado con éxito
Pulsa para continuar...
```

Ejecuciones malas:

El fichero de frecuencias esté creado con el nombre del usuario

```
1. Análisis de datos
=================================================================================================
En esta opción se pide 3 nombres:
        * Nombre del fichero en el que se encuentran las palabras o términos a buscar: (sword.txt)
        * Nombre del fichero donde se almacenan los correos electrónicos: (Emails.txt)
        * Nombre del fichero donde se desea almacenar el resultado del análisis
=================================================================================================
1 --> Nombre del fichero de palabras
Por favor, introduce el nombre del fichero de palabras: sword.txt
El fichero sword.txt existe.
==================================================
2 --> Nombre del fichero de correos
Por favor, introduce el nombre del fichero de correos: Emails.txt
El fichero  existe.
==================================================
3 --> Nombre del fichero para guardar los resultados
Por último, introduce el nombre del fichero resultados(solo el nombre): r1
El archivo r1.freq ya existe en la carpeta actual.
Pulsa para continuar...
```

Se introduzca un nombre de fichero de palabras o correos diferente al establecido

```
1. Análisis de datos
=================================================================================================
En esta opción se pide 3 nombres:
        * Nombre del fichero en el que se encuentran las palabras o términos a buscar: (sword.txt)
        * Nombre del fichero donde se almacenan los correos electrónicos: (Emails.txt)
        * Nombre del fichero donde se desea almacenar el resultado del análisis (sin extensión)
=================================================================================================
1 --> Nombre del fichero de palabras
Por favor, introduce el nombre del fichero de palabras: sword.txt
==================================================
2 --> Nombre del fichero de correos
Por favor, introduce el nombre del fichero de correos: Emails.txt
==================================================
3 --> Nombre del fichero para guardar los resultados
Por último, introduce el nombre del fichero resultados(solo el nombre): resultado
El archivo resultado.freq ya existe en la carpeta actual.
Pulsa para continuar...
```

No se introduzca nada

```
1. Análisis de datos
=================================================================================================
En esta opción se pide 3 nombres:
        * Nombre del fichero en el que se encuentran las palabras o términos a buscar: (sword.txt)
        * Nombre del fichero donde se almacenan los correos electrónicos: (Emails.txt)
        * Nombre del fichero donde se desea almacenar el resultado del análisis
=================================================================================================
1 --> Nombre del fichero de palabras
Por favor, introduce el nombre del fichero de palabras: sword.txt
El fichero sword.txt existe.
==================================================
2 --> Nombre del fichero de correos
Por favor, introduce el nombre del fichero de correos:
No has introducido nada. Vuelva a esta opción e ingresa un nombre válido
Pulsa para continuar...
```

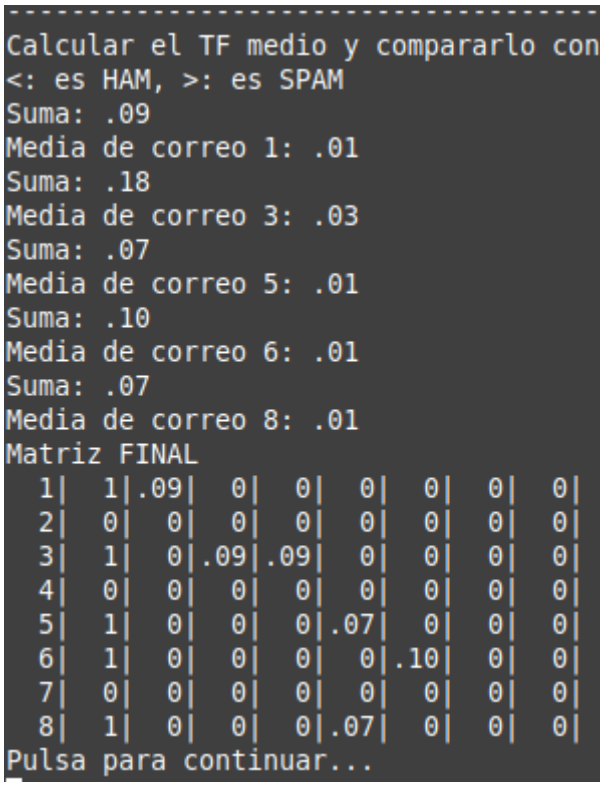
4.3 Opción 2

Ejecuciones buenas:

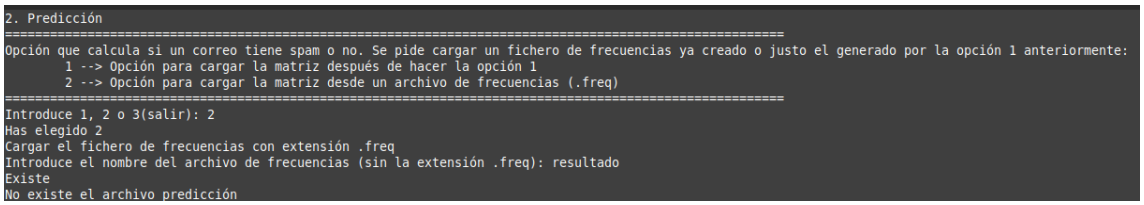
Se utiliza la opción 1 y posteriormente la opción 2, por lo que se hace la previsión sobre la matriz generada de la opción 1.

```
2. Predicción
=========================================================================================================
Opción que calcula si un correo tiene spam o no. Se pide cargar un fichero de frecuencias ya creado o justo el generado por la opción 1 anteriormente:
        1 --> Opción para cargar la matriz después de hacer la opción 1
        2 --> Opción para cargar la matriz desde un archivo de frecuencias (.freq)
=========================================================================================================
Introduce 1, 2 o 3(salir): 1
Has elegido 1
Hay que ver si se ha usado la opcion 1 previamente
Se ha utilzado la opcion 1: Existe el archivo creado anteriromente: resultado.freq
Numero de filas: 8
Numero de columnas: 9
Número de filas en la función: 8
Número de columnas en la función: 9
Nombre del fichero de frecuencias: resultado
Visualizar la matriz cargada
  1|  1|  1|  0|  0|  0|  0|  0| 10|
  2|  0|  0|  0|  0|  0|  0|  0| 11|
  3|  1|  0|  1|  1|  0|  0|  0|  9|
  4|  0|  0|  0|  0|  0|  0|  0|  6|
  5|  1|  0|  0|  0|  1|  0|  0|  8|
  6|  1|  0|  0|  0|  0|  1|  0|  8|
  7|  0|  0|  0|  0|  0|  0|  0|  9|
  8|  1|  0|  0|  0|  1|  0|  0|  8|
Total palabras correo 1: 10
Total palabras correo 2: 11
Total palabras correo 3: 9
Total palabras correo 4: 6
Total palabras correo 5: 8
Total palabras correo 6: 8
Total palabras correo 7: 9
Total palabras correo 8: 8
TFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
  1|  1|.10|  0|  0|  0|  0|  0|
  2|  0|  0|  0|  0|  0|  0|  0|
  3|  1|  0|.11|.11|  0|  0|  0|
  4|  0|  0|  0|  0|  0|  0|  0|
  5|  1|  0|  0|  0|.12|  0|  0|
  6|  1|  0|  0|  0|  0|.12|  0|
  7|  0|  0|  0|  0|  0|  0|  0|
  8|  1|  0|  0|  0|.12|  0|  0|
```

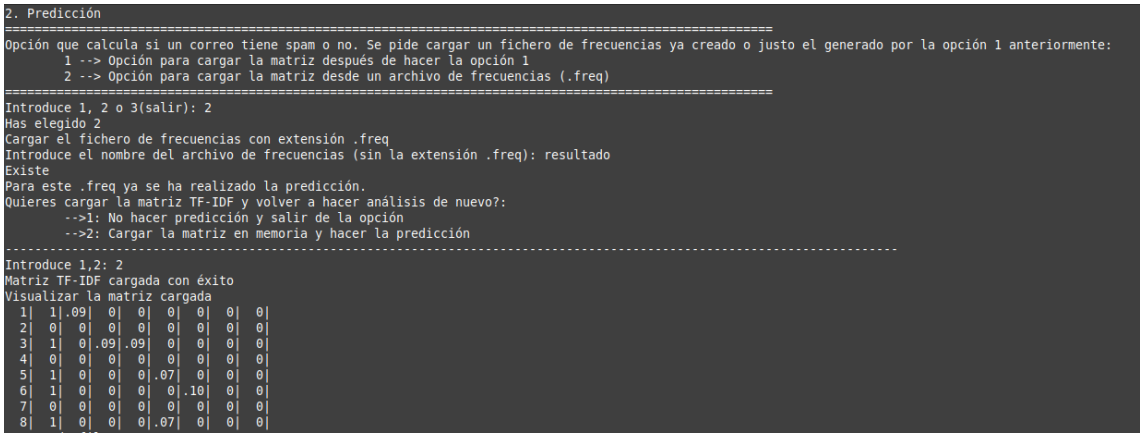
```
Calcular el IDF de cada término. Para ello, calcular para cada columna, todas las celdas que no sean 0.

IDF: log10 (nº total de correos/nº correos con el término x)
valor matriz = .10
Entra 1 veces
valor matriz = .11
Entra 2 veces
valor matriz = .11
Entra 3 veces
valor matriz = .12
Entra 4 veces
valor matriz = .12
Entra 5 veces
valor matriz = .12
Entra 6 veces
Contador IDF ya calculado
   |   |  1|  1|  1|  2|  1|  0|
-----------------------------------------------------------------------------
prueba
Resultado del logaritmo: .90
Resultado del logaritmo: .90
Resultado del logaritmo: .90
Resultado del logaritmo: .60
Resultado del logaritmo: .90
Visualizar la matriz con la fila final: IDF ya calculado
  1|  1|.10|  0|  0|  0|  0|  0|
  2|  0|  0|  0|  0|  0|  0|  0|
  3|  1|  0|.11|.11|  0|  0|  0|
  4|  0|  0|  0|  0|  0|  0|  0|
  5|  1|  0|  0|  0|.12|  0|  0|
  6|  1|  0|  0|  0|  0|.12|  0|
  7|  0|  0|  0|  0|  0|  0|  0|
  8|  1|  0|  0|  0|.12|  0|  0|
   |   |.90|.90|.90|.60|.90|  0|
-----------------------------------------------------------------------------
Ahora, ya toca calcular el TF x IDF de cada celda. Para ello hay que multiplicar el TF de cada celda por el IDF de la columna correspondiente, donde su valor está en la fila de abajo del todo.
Matriz con TF x IDF ya calculado para cada celda
  1|  1|.09|  0|  0|  0|  0|  0|
  2|  0|  0|  0|  0|  0|  0|  0|
  3|  1|  0|.09|.09|  0|  0|  0|
  4|  0|  0|  0|  0|  0|  0|  0|
  5|  1|  0|  0|  0|.07|  0|  0|
  6|  1|  0|  0|  0|  0|.10|  0|
  7|  0|  0|  0|  0|  0|  0|  0|
  8|  1|  0|  0|  0|.07|  0|  0|
```



Se introduce el fichero de frecuencias, se encuentra y no se ha realizado sobre éste, la previsión, por lo que se ejecuta la previsión.



Se introduce el fichero de frecuencias, se encuentra, pero se ha realizado sobre éste la previsión. Entonces el usuario, escoge la opción cargar la matriz TF-IDF y volver a hacerla



Ejecuciones malas

Se introduzca un 1 cuando el usuario no ha hecho el análisis previo

```
2. Predicción
================================================================================================
Opción que calcula si un correo tiene spam o no. Se pide cargar un fichero de frecuencias ya creado o justo el generado por la opción 1 anteriormente:
        1 --> Opción para cargar la matriz después de hacer la opción 1
        2 --> Opción para cargar la matriz desde un archivo de frecuencias (.freq)
================================================================================================
Introduce 1, 2 o 3(salir): 1
Has elegido 1
Hay que ver si se ha usado la opcion 1 previamente
No se ha usado todavía la opción análisis.
Pulsa para continuar...
```

Se introduzca un fichero de frecuencias que no exista

```
2. Predicción
================================================================================================
Opción que calcula si un correo tiene spam o no. Se pide cargar un fichero de frecuencias ya creado o justo el generado por la opción 1 anteriormente:
        1 --> Opción para cargar la matriz después de hacer la opción 1
        2 --> Opción para cargar la matriz desde un archivo de frecuencias (.freq)
================================================================================================
Introduce 1, 2 o 3(salir): 2
Has elegido 2
Cargar el fichero de frecuencias con extensión .freq
Introduce el nombre del archivo de frecuencias (sin la extensión .freq): r7
No existe el archivo r7.freq. Saliendo de la opción 2.
Pulsa para continuar...
```

Se introduzca un fichero de frecuencias que exista con su correspondiente fichero de predicción, y se decida no hacer nada.

```
2. Predicción
================================================================================================
Opción que calcula si un correo tiene spam o no. Se pide cargar un fichero de frecuencias ya creado o justo el generado por la opción 1 anteriormente:
        1 --> Opción para cargar la matriz después de hacer la opción 1
        2 --> Opción para cargar la matriz desde un archivo de frecuencias (.freq)
================================================================================================
Introduce 1, 2 o 3(salir): 2
Has elegido 2
Cargar el fichero de frecuencias con extensión .freq
Introduce el nombre del archivo de frecuencias (sin la extensión .freq): resultado
Existe
Para este .freq ya se ha realizado la predicción.
Quieres cargar la matriz TF-IDF y volver a hacer análisis de nuevo?:
        -->1: No hacer predicción y salir de la opción
        -->2: Cargar la matriz en memoria y hacer la predicción
------------------------------------------------------------------------------------------------------------------------
Introduce 1,2: 1
se ha introducido un 1. Saliendo de la opción...
Pulsa para continuar...
```

4.4 Opción 3

4.4.1. Opción 1 de la 3

Ejecución buena

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 1
---------------------------------------------------------------------------------------------
1- Informe de formato fila/columna de un término.
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
Existe el archivo resultado.freq

Término                                        Nºveces que aparece
-----------------------------------------------------------------------------
Oferta                                          1
Gratis                                          1
Gana                                            1
Descuento                                       2
Felicidades                                     1
Registro                                        0
Pulsa para continuar...
```

Ejecución mala

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 1
---------------------------------------------------------------------------------------------
1- Informe de formato fila/columna de un término.
Introduce el nombre del fichero de frecuencias(con la extensión): res.freq
El archivo res.freq no existe
Pulsa para continuar...
```

4.4.2. Opción 2 de la 3

Ejecución buena

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 2
---------------------------------------------------------------------------------------------
2- Informe de un término en particular
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
Introduce el término a buscar en minúsculas y sin caracteres especiales: descuento
Buscando descuento en el archivo Fraud_words.txt...

descuento encontrado en la posición: 4 de Fraud_words.txt

Posiciones: 5 | 8 |
Buscando en el archivo Emails.txt si aparece el término...

Está en 2 correos.

Índice del correo   Contenido del correo
-------------------------------------------------------------------------
5                    5|Descuento del 20% en tu próxima compra con noso
8                    8|Descuento del 20% en tu próxima compra con noso
Pulsa para continuar...
```

Ejecuciones malas

No exista el fichero de frecuencias

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 2
---------------------------------------------------------------------------------------------
2- Informe de un término en particular
Introduce el nombre del fichero de frecuencias(con la extensión): res.freq
res.freq no existe.
Pulsa para continuar...
```

No exista el término a buscar en el archivo de palabras

```
3. Informes de resultados
=============================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

=============================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
=============================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 2
---------------------------------------------------------------------------------------------
2- Informe de un término en particular
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
Introduce el término a buscar en minúsculas y sin caracteres especiales: prueba
Buscando prueba en el archivo Fraud_words.txt...
prueba no encontrado
Pulsa para continuar...
```

El término exista, pero no está en ningún correo

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 2
------------------------------------------------------------------------------------------------
2- Informe de un término en particular
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
Introduce el término a buscar en minúsculas y sin caracteres especiales: registro
Buscando registro en el archivo Fraud_words.txt...

registro encontrado en la posición: 6 de Fraud_words.txt

Posiciones:
Buscando en el archivo Emails.txt si aparece el término...
El término no aparece en ningún correo de Email.txt
Pulsa para continuar...
```

4.4.3. Opción 3 de la 3

Ejecución buena

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 3
------------------------------------------------------------------------------------------------
3- Informe número de términos en un correo
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
resultado.freq existe.
Introduce el identificador del correo: 3
Buscando 3 en el archivo en la matriz...

El identificador pertenece a un correo. Buscando las posiciones de los términos...

Posiciones: 2 | 3 |
Hay un total de 2 palabras de spam en el correo con id: 3

--> Los términos son: Gratis  |  Gana  |
Pulsa para continuar...
```

Ejecución mala

No exista el archivo de frecuencias

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 3
Opción 3 escogida
3- Informe número de términos en un correo
Introduce el nombre del fichero de frecuencias(con la extensión): r4.freq
r4.freq no existe.
Pulsa para continuar...
```

El id del correo no existe

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 3
------------------------------------------------------------------------------------------------
3- Informe número de términos en un correo
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
resultado.freq existe.
Introduce el identificador del correo: 10
Buscando 10 en el archivo en la matriz...

El 10 no pertenece a ningún identificador de correo.
Pulsa para continuar...
```

El id introducido es negativo

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 3
------------------------------------------------------------------------------------------------
3- Informe número de términos en un correo
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
resultado.freq existe.
Introduce el identificador del correo: -10
El id no puede ser negativo
Pulsa para continuar...
```

El correo no tenga ningún término:

```
3. Informes de resultados
================================================================================================
En esta opción tienes 3 tipos diferentes de informe y dependiendo del número pulsado, escoges una. Las opciones son:

        1--> Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del
conjunto de datos analizado aparece
        2--> Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos
donde aparece
        3-->* Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.

================================================================================================
En las 3 opciones se va a pedir un fichero de frecuencias para poder realizar los informes correspondientes. Aparte, para
no tener que pedir otra vez los ficheros de palabras y email, se usarán los que tienen el nombre predeterminado de la práctica:
       --> Email: Emails.txt
       --> Términos --> Fraud_words.txt
================================================================================================
Introduce un número (1, 2, 3 o 4 (salir)): 3
------------------------------------------------------------------------------------------------
3- Informe número de términos en un correo
Introduce el nombre del fichero de frecuencias(con la extensión): resultado.freq
resultado.freq existe.
Introduce el identificador del correo: 2
Buscando 2 en el archivo en la matriz...

El identificador pertenece a un correo. Buscando las posiciones de los términos...

No hay ninguna palabra de spam en el correo con id: 2
Pulsa para continuar...
```

4.5: Opción 4

5. Bibliografía

Para poder realizar la práctica, he utilizado en su mayoría, la documentación de la asignatura, pero también he necesitado buscar algo de información en Internet. Estas son las fuentes:

Hacer el filtro de quitar símbolos especiales, pasar a minúsculas y borrar números:

Anexo 2 del pdf de la práctica.

Logaritmos:

https://www.linuxquestions.org/questions/programming-9/calculate-logarithm-in-bash-script-690036/

Números flotantes:

https://www.youtube.com/watch?v=vTTV2cUkkEU&t=920s

Introducir solo números:

https://www.lawebdelprogramador.com/codigo/Linux-Unix-Shell-Scripting/3177-bash-linux-introducir-solo-numeros.html

Comando awk (búsqueda en archivos):

https://geekland.eu/uso-del-comando-awk-en-linux-y-unix-con-ejemplos/

Matrices y arrays en bash:

https://geekflare.com/es/bash-arrays/