



Práctica 1

Estado	Done
Asignatura	 Sistemas Operativos
Fecha de vencimiento	@27 de octubre de 2023
Tipo	Práctica
Alumno	 Setrakian Bearzotti, Pablo
Titulación	Ingeniería Informática en Tecnologías de la Información
Curso	3º
Grupo	Lunes de 08:00 a 10:00
Convocatoria	Ordinaria

1. Memoria explicativa

1.2 Introducción al supuesto práctico

- 1.2.1 ¿Qué es el TF-IDF?
- 1.2.2 Propósito de la práctica

2. Trayectoria

2. Manual de usuario

- 2.1 Análisis
- 2.2 Predicción
- 2.3 Informes
- 2.4 Ayuda

3. Manual de programador

- 2.1 Menú principal
- 2.2 Análisis
- 2.3 Predicción
- 2.4 Informes

4. Juego de pruebas

5. Bibliografía

1. Memoria explicativa

1.2 Introducción al supuesto práctico

1.2.1 ¿Qué es el TF-IDF?

TF-IDF (Term Frequency-Inverse Document Frequency) es una medida numérica que evalúa la relevancia de una palabra en un documento dentro de una colección de documentos. Se utiliza en la recuperación de información y la minería de texto. El valor TF-IDF aumenta con la frecuencia de una palabra en un documento, pero se equilibra teniendo en cuenta la frecuencia de la palabra en la colección de documentos, lo que ayuda a lidar con palabras comunes.

El cálculo de TF-IDF se basa en dos medidas: la frecuencia de término (TF), que mide cuántas veces aparece una palabra en un documento, y la frecuencia inversa de documento (IDF), que mide si una palabra es común o no en la colección. Diversas fórmulas pueden calcular TF, como la frecuencia bruta, binaria, logarítmica o normalizada. IDF se obtiene dividiendo el número total de documentos entre los que contienen el término y tomando el logaritmo del resultado. Luego, el valor TF-IDF se calcula multiplicando TF e IDF.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Si el numero total de expresiones en el documento es 0 el TF no se puede calcular
(intedeterminación)

$$IDF = \log\left(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}}\right)$$

Si el término no está en la colección se producirá una división-por-cero. Por lo tanto, es común ajustar esta fórmula a $1 + |d \in D : t \in d|$

$$TF-IDF = TF * IDF$$

1.2.2 Propósito de la práctica

Para esta práctica he tenido que crear una aplicación que analiza el contenido de un conjunto de correos electrónicos con el fin de identificar posibles amenazas. Para llevar a cabo este proceso, se utilizan dos archivos: "sword.txt", que contiene palabras y frases comúnmente asociadas con correos no deseados, y "emails.txt", que incluye el texto de aproximadamente 12,000 correos electrónicos (el cual he reducido a 10), algunos de los cuales son peligrosos, masivos o normales.

La aplicación calcula cuántas veces aparece cada palabra del archivo "sword.txt" en todos los correos electrónicos. Con estos resultados, se determina si un correo electrónico es considerado potencialmente peligroso.

La aplicación ofrece un menú que permitirá acceder a todas las funciones. Estas opciones son:

1. Análisis de datos:

- El usuario proporciona los nombres de los archivos de palabras a buscar, correos electrónicos y el archivo para almacenar los resultados del análisis.
- Realiza un análisis sobre una matriz con tantas filas como correos electrónicos y tantas columnas como expresiones sospechosas (como mínimo).
- Guarda los resultados en un archivo con extensión ".freq".

2. Predicción:

- Permite cargar un análisis de frecuencias previo o crear uno nuevo..
- Calcula la métrica TF-IDF para predecir si un correo es spam o ham.
- Guarda los resultados en un archivo con extensión ".tfidf".

3. Informes de resultados:

- Genera diferentes informes en formato de tabla:
 - Informe que muestra cuántas veces aparece cada término en los correos electrónicos.
 - Informe que muestra los correos en los que aparece un término específico.
 - Informe que muestra cuántos términos analizados aparecen en un correo específico.

4. Ayuda:

- Proporciona información de ayuda sobre el funcionamiento de la aplicación.

5. Salir:

- Finaliza la aplicación.

2. Trayectoria

En primer lugar, antes de que comenzase a codificar el programa final, hice muchos scripts donde intenté implementar cada una de las distintas opciones y las distintas funcionalidades que implica cada opción por separado, utilizando multiples ficheros de prueba.

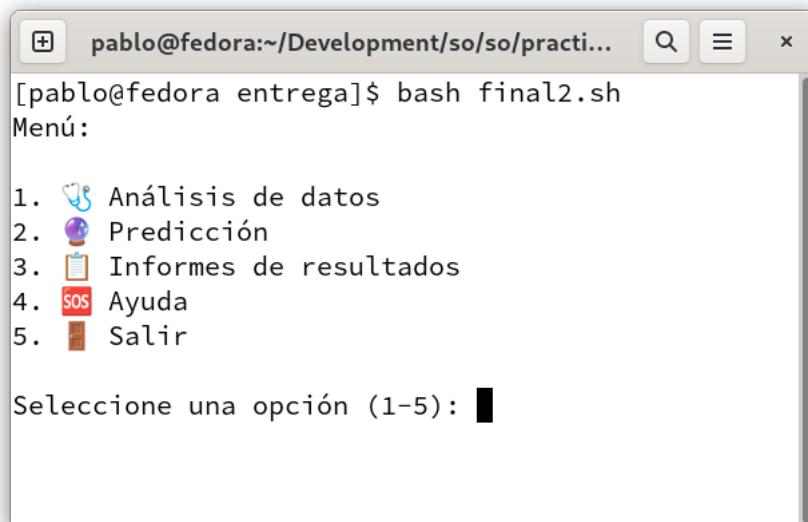
Una vez empecé a adquirir mayor entendimiento del lenguaje Bash, comencé a desarrollar la aplicación.

El principal problema al que me he enfrentado con esta practica ha sido lograr entender lo que debia de hacer el programa en cada una de las distintas opciones.

Por otra parte la falta de optimización y repetición de codigo ha sido otro de mis obstaculos, el cual he podido resolver en gran medida desarrollando funciones.

2. Manual de usuario

Para poder ejecutar la aplicación, el usuario debe acceder al directorio donde se encuentre el script y ejecutar el siguiente comando en el terminal: `bash final2.sh`.



2.1 Análisis

Para realizar en análisis de datos, el usuario debe seleccionar la opción 1 y pulsar la tecla enter. Una vez dentro de la opción se le preguntará por el nombre de distintos ficheros, estos son los ficheros que contienen los correos electrónicos, las palabras sospechosas y el fichero que contendrá el análisis de frecuencias.

Utilizaré ficheros de ejemplo de tamaño reducido para esta explicación de manera que estos sean facilmente legibles para el usuario.

- emails.txt

```
1|Supply Quality China's EXCLUSIVE dimensions at Unbeatable Price.Dear Sir, We are pleased to inform you as one of China's largest
2|over. SidLet me know. Thx.|0|
3|Dear Friend,Greetings to you.I wish to accost you with a request that would be of immense benefit to both of us. Being an execut
4|MR. CHEUNG PUIHANG SENG BANK LTD.DES VOEUX RD. BRANCH,CENTRAL HONG KONG,HONG KONG. Let me start by introducing myself. I am Mr.C
5|Not a surprising assessment from Embassy.|0|
6|"Monica -Huma Abedin <Huma@clintonemail.com>Tuesday June 29 2010 6:01 AM'hanleymr@state.gov'; HRe:is already is locked for tonit
7|Pis print.H <hrod17@clintonemail.com>Thursday October 8 2009 8:01 PM'JilolyLC@state.gov'Fw: WHI - powder coatingB6|0|
8|"Dear Tom--H <hrod17@clintonemail.com>Friday December 11 2009 5:41 PMCould we schedule a call?I have several matters I'd like to
9|Greetings from barrister Robert Williams=2CDear friend=2C I know that my letter will come to you as a surprise=2C b=aed on the
10|FYI. Thanks again for signing the book ---- and I do hope you get royalties from Mongolia! Also thought you would like to see th
```

- sword.txt

```
4U
Accept credit cards
Act now! Don't hesitate!
Additional income
Addresses on CD
```

All natural
Amazing
Apply Online
As seen on
Auto email removal
Avoid bankruptcy
Be amazed
Be your own boss
Being a member
Big bucks
Bill
Billing address
Billion dollars
Brand new pager
Bulk email
Buy direct
Buying judgments
Cable converter
Call free
Call now
Calling creditors
Can't live without
Cancel at any time
Cannot be combined with any other offer
Cash bonus
Cashcashcash
Casino
Cell phone cancer scam
Cents on the dollar
Check or money order
Claims not to be selling anything
Claims to be in accordance with some spam law
Claims to be legal
Claims you are a winner
Click below
Click here link
Click to remove
Click to remove mailto
Compare rates
Compete for your business
Confidentially on all orders
Congratulations
Consolidate debt and credit
Copy accurately
Copy DVDs
Credit bureaus
Credit card offers
Cures baldness
Dear email
Dear friend
Dear somebody
Different reply to
Dig up dirt on friends
Direct email
Direct marketing
Discusses search engine listings
Do it today
Don't delete
Drastically reduced
Earn per week
Easy terms
Eliminate bad credit
Email harvest
Email marketing
Expect to earn
Fantastic deal
Fast Viagra delivery
Financial freedom
Find out anything
For free
For instant access
For just \$
Free access
Free cell phone
Free consultation
Free DVD
Free grant money
Free hosting
Free installation
Free investment
Free leads
Free membership
Free money
Free offer

Free preview
Free priority mail
Free quote
Free sample
Free trial
Free website
Free transfer
Full refund
Get It Now
Get paid
Get started now
Gift certificate
Great offer
Guarantee
Have you been turned down?
Hidden assets
Home employment
Human growth hormone
If only it were that easy
In accordance with laws
Increase sales
Increase traffic
Insurance
Income
Investment decision
It's effective
Join millions of Americans
Limited time only
Long distance phone offer
Lose weight spam
Lower interest rates
Lower monthly payment
Lowest price
Luxury car
Mail in order form
Marketing solutions
Mass email
Meet singles
Member stuff
Message contains disclaimer
MLM
Million in cash
Money back
Money making
Month trial offer
More Internet traffic
Mortgage rates
Multi level marketing
Name brand
New customers only
New domain extensions
Nigerian
No age restrictions
No catch
No claim forms
No cost
No credit check
No disappointment
No experience
No fees
No gimmick
No inventory
No investment
No medical exams
No middleman
No obligation
No purchase necessary
No questions asked
No selling
No strings attached
Not intended
Off shore
Offer expires
Offers coupon
Offers extra cash
Offers free
Once in lifetime
One hundred percent free
One hundred percent guaranteed
One time mailing
Online biz opportunity
Online pharmacy
Only \$
Opportunity

Opt in
Order now
Order status
Orders shipped by priority mail
Outstanding values
Pennies a day
People just leave money laying around
Please read
Potential earnings
Print form signature
Print out and fax
Produced and sent out
Profits
Promise you
Pure profit
Real thing
Refinance home
Removal instructions
Remove
Remove subject
Removes wrinkles
Reply remove subject
Requires initial investment
Reserves the right
Reverses aging
Risk free
Round the world
Safeguard notice

Cualquier error al introducir el nombre de un fichero le será notificado al usuario a través de un mensaje.

pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh

[pablo@fedora entrega]\$ bash final2.sh

Menú:

1. 📈 Análisis de datos
2. 🎮 Predicción
3. 📊 Informes de resultados
4. 🚨 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 1

✉️ Introduzca el nombre del fichero que contiene los correos electrónicos:
🚩 Entrada inválida. Le quedan 2 intentos.

✉️ Introduzca el nombre del fichero que contiene los correos electrónicos: noexiste.txt
🚩 El fichero no existe. Le quedan 1 intentos.

✉️ Introduzca el nombre del fichero que contiene los correos electrónicos: sword.txt
🚫 El contenido del fichero no sigue la estructura requerida. Le quedan 0 intentos.

Introduzca cualquier tecla para regresar al menú: █

Es necesario que el fichero que contiene los correos electrónicos tenga la siguiente estructura en cada linea:

ID|Contenido del correo|Resultado|

donde:

- ID debe de ser un número entero
- Contenido del correo puede ser cualquier secuencia de caracteres
- Resultado debe ser bien el número entero 0 o el número entero 1.

Ejemplo: 3|Gana un viaje gratis a destinos exóticos. ¡Regístrate ahora!|1|

Despues de haber introducido los ficheros que contienen los correos electrónicos y las palabras sospechosas, se le pedira al usuario si desea ademas crear un fichero con extensión .freq donde se guardara el analisis.

Si el usuario solo desea realizar un analisis en la ejecución para posteriormente acceder a los informes no tiene porque crear el fichero de frecuencias.

```
pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh
Menú:

1. 📈 Análisis de datos
2. 🌐 Predicción
3. 📄 Informes de resultados
4. 🚨 Ayuda
5. 🛡 Salir

Seleccione una opción (1-5): 1
☒ Introduzca el nombre del fichero que contiene los correos electrónicos: emails.txt
☒ Introduzca el nombre del fichero que contiene las expresiones sospechosas: sword.txt
☒ ¿Desea crear un fichero .freq donde se escribirá el análisis de los correos electrónicos?
    Si posteriormente desea realizar una predicción este fichero es necesario
    Si por el contrario solo desea acceder a los informes puede prescindir del fichero
Introduzca una opción (s/n): n
██████████ | \ 100% [ ✓ Análisis completado con éxito ]
Introduzca cualquier tecla para regresar al menú: █
```

Si por el contrario el usuario desea realizar una predicción deberá crear el fichero.

- analisis.freq

En la primera columna se almacena el identificador del correo electrónico, en la segunda columna el numero de palabras que contiene y en la tercera el valor de spam o ham. 1 representa spam, 0 representa ham. En el resto de columnas se almacenan las coincidencias por cada expresión del fichero de palabras.

2.2 Predicción

Para realizar en análisis de datos, el usuario debe seleccionar la opción 2 y pulsar la tecla enter.

Si el usuario acaba de realizar el análisis de datos y no seleccionó crear un fichero de frecuencias se le avisará de que no se puede crear una predicción en base al análisis que acaba de hacer, pero que puede cargar un fichero de frecuencias para una nueva predicción si así lo desea.

```

pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh
Menú:

1. 📈 Análisis de datos
2. 🎯 Predicción
3. 📄 Informes de resultados
4. 💊 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 2
🌐 ⚠ Se ha encontrado un análisis en la ejecución pero no existe fichero de frecuencias (.freq).
Vuelva a realizar el análisis y asegúrese de crear el fichero de frecuencias (.freq)
¿Desea cargar un fichero de frecuencias para realizar una predicción? (s/n): 

```

Si por el contrario el usuario eligió crear un fichero de frecuencias, el fichero le avisará de que existe un análisis en la ejecución del programa y que si desea utilizarlo o desea cargar otro distinto.

La razón por la que se necesita un fichero de frecuencias aunque se utilice el análisis en ejecución es porque el fichero resultante de la predicción (.tfidf) debe tener el mismo nombre que el fichero de frecuencias (.freq).

Veamos un ejemplo de una predicción utilizando el análisis encontrado en la ejecución del programa:

```

pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh
Menú:

1. 📈 Análisis de datos
2. 🎯 Predicción
3. 📄 Informes de resultados
4. 💊 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 2
🌐 Se ha encontrado un análisis en la ejecución. ¿Desea utilizarlo? (s/n): s
load_tfidf= 0
use_current_analysis= 1
[REDACTED] | \ 100% [ ✅ TF-IDF calculado con éxito ] 8 ]
[REDACTED] | \ 100% [ ✅ Predicción calculada con éxito ] 8 ]
*****
* Matriz TF-IDF guardada en: /home/pablo/Development/so/so/practica1/entrega)/análisis.tfidf *
***** 

Introduzca cualquier tecla para regresar al menú: 

```

En este caso se calcula la matriz con la métrica TF-IDF y posteriormente una predicción en base a los resultados de esta métrica. El resultado de la predicción para cada correo (0 o 1) se guardará en una nueva fila de la matriz.

En el caso de que existiera algún correo electrónico vacío, este no se tendría en cuenta para el cálculo del TF-IDF y la línea correspondiente no se utilizaría en la matriz.

- emails2.txt:

```

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|¡Felicitaciones! Eres el ganador de nuestro concurso mensual|1|
7|Actualización de política de privacidad y términos de servicio|0|
8|                                |0|
9|Descuento del 20% en tu próxima compra con nosotros|1|

```

- c.freq

```

1:10:1:1:0:0:0:0:0:
2:11:0:0:0:0:0:0:0:
3:9:1:0:1:1:0:0:0:
4:6:0:0:0:0:0:0:0:
5:8:1:0:0:0:1:0:0:
6:8:1:0:0:0:0:1:0:
7:9:0:0:0:0:0:0:0:
8:0:0:0:0:0:0:0:0: Linea vacia, no contiene terminos
9:8:1:0:0:0:1:0:0:

```

```

pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh
Menú:

1. 📈 Análisis de datos
2. 🌐 Predicción
3. 📄 Informes de resultados
4. 💬 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 2
🔍 Se ha encontrado un análisis en la ejecución. ¿Desea utilizarlo? (s/n): s
⚠️ Advertencia: El E-Mail 8 esta vacio, no se tendra en cuenta en cuenta para el calculo del TF-IDF.
load_tfidf= 0
use_current_analysis= 1
[██████████] \ 100% [ ✓ TF-IDF calculado con éxito ] 9 ]
[██████████] \ 100% [ ✓ Predicción calculada con éxito ] 9 ]
*****
* 📁 Matriz TF-IDF guardada en: /home/pablo/Development/so/so/practica1/entrega)/c.tfidf *
*****


Introduzca cualquier tecla para regresar al menú:█

```

- c.tfidf

```

1:10:1:.0950:0:0:0:0:0:
2:11:0:0:0:0:0:0:0:
3:9:1:0:.1045:.1045:0:0:0:
4:6:0:0:0:0:0:0:0:0:
5:8:1:0:0:0:.0780:0:0:0:
6:8:1:0:0:0:0:.1140:0:0:
7:9:0:0:0:0:0:0:0:0:      El E-Mail 8 no aparece
9:8:1:0:0:0:.0780:0:0:0:

```

Tambien existe la posibilidad de cargar un fichero de frecuencias, para realizar la predicción. Si el fichero que deseamos cargar no sigue una estructura valid, se mostrará un mensaje avisando de ello.

Por ejemplo:

- c.freq

```

1:10:1:1:0:0:0:0:0:
2:11:0:0:0:0:0:0:0:
3:9:1:0:1:1:0:0:0:
4:6:0:0:0:0:0:0:0:
5:8:1:0:0:0:1:0:0:

```

```
6:8:1:0:0:0:0:1:0:  
7:9:0:0:0:0:0:0:0:  
Hola me llamo pablo  
9:8:1:0:0:0:1:0:0:
```

The screenshot shows a terminal window titled "pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh". The window contains the following text:

```
Menú:  
1. 📈 Análisis de datos  
2. 🌎 Predicción  
3. 📄 Informes de resultados  
4. 🚨 Ayuda  
5. 🚪 Salir  
Seleccione una opción (1-5): 2  
🌐 Se ha encontrado un análisis en la ejecución. ¿Desea utilizarlo? (s/n): n  
Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: c.freq  
🚩 El fichero no sigue un formato valido. Le quedan 2 intentos.  
Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: █
```

En el caso de que no exista ningun problema con el fichero, se calculará el TF-IDF de manera común.

Por otra parte, si el programa encuentra un fichero con extensión ".tfidf" asociando al fichero con extension ".freq" introducido, se informaría al usuario que el fichero con la métrica TF_idf también se ha realizado con anterioridad y se pide conformidad para cargar todos los datos en las matrices correspondientes.

Si el usuario elige cargar el fichero ".tfidf" asociado, este se escribirá al final del fichero existente, lo cual es muy util para una comparación de resultados en el caso de que el requerimiento para que un correo fuera considerado "ham" o "spam" cambiase

```
pablo@fedora:~/Development/so/so/practical/entrega — bash final2.sh
Menú:

1. 📈 Análisis de datos
2. 🎠 Predicción
3. 📊 Informes de resultados
4. 📡 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 2
💡 Se ha encontrado un análisis en la ejecución. ¿Desea utilizarlo? (s/n): n
Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: c.freq
⚠️ El fichero no sigue un formato valido. Le quedan 2 intentos.
Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: analysis.freq
Se ha encontrado un fichero con extensión .tfidf para el fichero .freq introducido, ¿Desea cargarlo para una nueva predicción?
s ⚡ El fichero con la matriz TF-IDF se utilizará para calcular una nueva predicción
n ⚡ El fichero con la matriz TF-IDF se borrará y se volverá a realizar el calculo del TF-IDF y la predicción
Seleccione una opción (s/n): s
load_tfidf= 1
use_current_analysis= 0
[██████████] / 100% [ ✅ TF-IDF cargado con éxito ] 10 ]
[██████████] / 100% [ ✅ Predicción calculada con éxito ] 10 ]
*****
* 📁 Matriz TF-IDF guardada en: /home/pablo/Development/so/so/practical/entrega/analysis.tfidf *
*****


Introduzca cualquier tecla para regresar al menú: █
```

- Entrada: análisis.freq
 - Salida: análisis.tfidf

En el caso de que el usuario elija no cargar el TF-IDF el fichero con extension ".tfidf" se borrará y se volverá a calcular desde cero el TF-IDF y por consiguiente la predicción y se reescribirá este nuevo resultado.

Tambien puede ocurrir que exista un fichero con extension ".tfidf" el cual no siga el formato correcto, en este caso se le avisará al usuario para que le eche un vistazo y lo corrija.

Ejemplo

- analisis.tfidf

```
pablo@fedora:~/Development/so/so/practica1/entrega— bash final2.sh
Menú:

1. 📈 Análisis de datos
2. 🎠 Predicción
3. 📄 Informes de resultados
4. 💬 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 2
🌐 Se ha encontrado un análisis en la ejecución. ¿Desea utilizarlo? (s/n): n
Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: analysis.freq
🔴 Advertencia: Se ha encontrado un fichero .tfidf no valido para el fichero .freq introducido.
    Por favor, revise el fichero en en el caso de que quiera utilizarlo para una nueva predicción.
Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: █
```

2.3 Informes

Para acceder a los informes, el usuario primero debe de haber realizado un análisis, de lo contrario el programa no le permitirá acceder a dicha opción.

```
pablo@fedora ~]$ bash final2.sh
[pablo@fedora entrega]$ bash final2.sh
Menú:
1. 📈 Análisis de datos
2. 🌎 Predicción
3. 📁 Informes de resultados
4. 📡 Ayuda
5. 🚪 Salir

Seleccione una opción (1-5): 3
🚫 Error: No se puede acceder a los informes sin realizar un análisis previo.
Introduzca cualquier tecla para regresar al menú: █
```

En el caso de que el usuario haya realizado un análisis previo la aplicación si le permitirá acceder a los informes.

Una vez dentro del menú de informes el usuario puede elegir entre distintas opciones.

- Opción 1:

```

Informes:
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. Del correo electrónico sólo se mostrarán los 50 primeros caracteres.
3. Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.
4. Regresar al menu principal

Selecciona una opción (1-4): 1
4U 0
Accept credit cards 0
Act now! Don't hesitate! 0
Additional income 0
Addresses on CD 0
All natural 0
Amazing 0
Apply Online 0
As seen on 0
Auto email removal 0
Avoid bankruptcy 0
Be amazed 0
Be your own boss 0
Being a member 0
Big bucks 0
Bill 0
Billing address 0
Billion dollars 0
Brand new pager 0
Bulk email 0
Buy direct 0
Buying judgments 0
Cable converter 0
Call free 0
Call now 0
Calling creditors 0
Can't live without 0
Cancel at any time 0
Cannot be combined with any other offer 0
Cash bonus 0
Cashcashcash 0
Casino 0
Cell phone cancer scam 0
Cents on the dollar 0
Check or money order 0
Claims not to be selling anything 0
Claims to be in accordance with some spam law 0
Claims to be legal 0
Claims you are a winner 0
Click below 0
Click here link 0
Click to remove 0
Click to remove mailto 0
Compare rates 0

```

Entrada:

```
1
```

Salida:

```

4U 0
Accept credit cards 0
Act now! Don't hesitate! 0
Additional income 0
Addresses on CD 0
All natural 0
Amazing 0
Apply Online 0
As seen on 0
Auto email removal 0
Avoid bankruptcy 0
Be amazed 0
Be your own boss 0
Being a member 0
Big bucks 0
Bill 0
Billing address 0
Billion dollars 0
Brand new pager 0
Bulk email 0
Buy direct 0
Buying judgments 0
Cable converter 0
Call free 0
Call now 0
Calling creditors 0
Can't live without 0
Cancel at any time 0
Cannot be combined with any other offer 0
Cash bonus 0
Cashcashcash 0
Casino 0
Cell phone cancer scam 0
Cents on the dollar 0
Check or money order 0
Claims not to be selling anything 0
Claims to be in accordance with some spam law 0
Claims to be legal 0
Claims you are a winner 0
Click below 0
Click here link 0
Click to remove 0
Click to remove mailto 0
Compare rates 0

```

Compete for your business 0
Confidentially on all orders 0
Congratulations 0
Consolidate debt and credit 0
Copy accurately 0
Copy DVDs 0
Credit bureaus 0
Credit card offers 0
Cures baldness 0
Dear email 0
Dear friend 1
Dear somebody 0
Different reply to 0
Dig up dirt on friends 0
Direct email 0
Direct marketing 0
Discusses search engine listings 0
Do it today 0
Don't delete 0
Drastically reduced 0
Earn per week 0
Easy terms 0
Eliminate bad credit 0
Email harvest 0
Email marketing 0
Expect to earn 0
Fantastic deal 0
Fast Viagra delivery 0
Financial freedom 0
Find out anything 0
For free 0
For instant access 0
For just \$ 0
Free access 0
Free cell phone 0
Free consultation 0
Free DVD 0
Free grant money 0
Free hosting 0
Free installation 0
Free investment 0
Free leads 0
Free membership 0
Free money 0
Free offer 0
Free preview 0
Free priority mail 0
Free quote 0
Free sample 0
Free trial 0
Free website 0
Free transfer 0
Full refund 0
Get It Now 0
Get paid 0
Get started now 0
Gift certificate 0
Great offer 0
Guarantee 0
Have you been turned down? 0
Hidden assets 0
Home employment 0
Human growth hormone 0
If only it were that easy 0
In accordance with laws 0
Increase sales 0
Increase traffic 0
Insurance 0
Income 0
Investment decision 0
It's effective 0
Join millions of Americans 0
Limited time only 0
Long distance phone offer 0
Lose weight spam 0
Lower interest rates 0
Lower monthly payment 0
Lowest price 0
Luxury car 0
Mail in order form 0
Marketing solutions 0
Mass email 0
Meet singles 0
Member stuff 0

Message contains disclaimer 0
MLM 0
Million in cash 0
Money back 0
Money making 0
Month trial offer 0
More Internet traffic 0
Mortgage rates 0
Multi level marketing 0
Name brand 0
New customers only 0
New domain extensions 0
Nigerian 0
No age restrictions 0
No catch 0
No claim forms 0
No cost 0
No credit check 0
No disappointment 0
No experience 0
No fees 0
No gimmick 0
No inventory 0
No investment 0
No medical exams 0
No middleman 0
No obligation 0
No purchase necessary 0
No questions asked 0
No selling 0
No strings attached 0
Not intended 0
Off shore 0
Offer expires 0
Offers coupon 0
Offers extra cash 0
Offers free 0
Once in lifetime 0
One hundred percent free 0
One hundred percent guaranteed 0
One time mailing 0
Online biz opportunity 0
Online pharmacy 0
Only \$ 0
Opportunity 1
Opt in 0
Order now 0
Order status 0
Orders shipped by priority mail 0
Outstanding values 0
Pennies a day 0
People just leave money laying around 0
Please read 0
Potential earnings 0
Print form signature 0
Print out and fax 0
Produced and sent out 0
Profits 0
Promise you 0
Pure profit 0
Real thing 0
Refinance home 0
Removal instructions 0
Remove 0
Remove subject 0
Removes wrinkles 0
Reply remove subject 0
Requires initial investment 0
Reserves the right 0
Reverses aging 0
Risk free 1
Round the world 0

- Opción 2:

```
pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh
Informes:
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. Del correo electrónico sólo se mostrarán los 50 primeros caracteres.
3. Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.
4. Regresar al menú principal

Seleccione una opción (1-4): 2
Introduzca una expresión: Risk free
9|Greetings from barrister Robert Williams=2CDear
Introduzca cualquier tecla para regresar al menú de informes:
```

Entrada:

```
Risk free
```

El correo electrónico completo es:

```
9|Greetings from barrister Robert Williams=2CDear friend=2C I know that my letter will come to you as a surprise=2C b=aed on the
```

- Opcion 3:

```
pablo@fedora:~/Development/so/so/practica1/entrega — bash final2.sh
Informes:
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. Del correo electrónico sólo se mostrarán los 50 primeros caracteres.
3. Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen.
4. Regresar al menú principal

Seleccione una opción (1-4): 3
Introduzca un identificador: 9
En el correo electrónico 9 aparecen 2 de un total de 200 expresiones sospechosas.

Introduzca cualquier tecla para regresar al menú de informes:
```

Entrada

```
9
```

Salida:

```
En el correo electrónico 9 aparecen 2 de un total de 200 expresiones sospechosas.
```

2.4 Ayuda

Si el usuario desea obtener ayuda sobre el funcionamiento de la aplicación deberá seleccionar la opción 4.

```
[pablo@fedora entrega]$ bash final2.sh
Menú:
1. Análisis de datos
2. Predicción
3. Informes de resultados
4. Ayuda
5. Salir

Seleccione una opción (1-5): 4
*****
* Ayuda *
*****

Esta aplicación realiza análisis de correos electrónicos para identificar spam.
Para ello utiliza la métrica TF-IDF (Term Frequency-Inverse Document Frequency), una medida numérica utilizada para evaluar la relevancia de palabras
en un documento en relación con una colección de documentos. Se emplea en la recuperación de información y la minería de texto, especialmente en motores
de búsqueda. La idea es que el valor tf-idf aumenta con la frecuencia de una palabra en un documento pero disminuye si la palabra es común en la colección,
lo que permite priorizar palabras más significativas.

Opciones disponibles:
1. Análisis de datos: Realiza el análisis de frecuencia de palabras en los correos.
2. Predicción: Calcula la métrica TF-IDF y predice si un correo es spam o ham.
3. Informes de resultados: Genera informes basados en los datos analizados.
4. Ayuda: Muestra esta información de ayuda.
5. Salir: Finaliza la aplicación.

Recuerda controlar los errores y seguir las instrucciones en cada opción.

Introduzca cualquier tecla para regresar al menú principal:
```

3. Manual de programador

2.1 Menú principal

El primer paso fue implementar un bucle `while`, para la creación del menu principal. Aquí se le pide al usuario que elija entre las diferentes opciones.

```
while [ true ]; do
    echo "Menú:"
    echo
    echo "1. Análisis de datos"
    echo "2. Predicción"
    echo "3. Informes de resultados"
    echo "4. Ayuda"
    echo "5. Salir"
    echo
    read -rp "Seleccione una opción (1-5): " choice
    case $choice in
        1)
            # Implementación del análisis
            ;;
        2)
            # Implementación de la predicción
            ;;
        3)
            # Implementación de los informes
            ;;
        4)
            # Mostrar ayuda
            ;;
        5)
            # Salir
            exit 0
            ;;
        -z)
            continue
            ;;
        esac
        clear
    done
```

Una vez conseguido el menu principal, comencé con la implementación de la opción 1.

2.2 Análisis

En primer lugar escribí una función que me permitiese dar el formato correcto al texto, puesto que iba a ser llamada multiples veces.

```
clean_text() {
    input="$1"
    cleaned_text=$(echo "$input" | awk '{print tolower($0)}' | awk '{ gsub(/[^[:alnum:]]/, " "); gsub(/ +/, " "); gsub(/\<[0-9]+\>/, "
    ) }'
}
```

No tuve el mismo afán de optimización al codificar el resto de la opción y termine con un código en el que existía demasiada redundancia a la hora de pedir datos al usuario, es por eso que opté por crear las siguientes funciones:

```
# Verifica si el usuario introduce un fichero correcto
verify_input_file() {
    file_var="$1"
    local prompt="$2"
    local check_exists="$3"
    local extension="$4"
    local structure_regex="$5"

    local i=3
    while [ "$i" -gt 0 ]; do
        read -rp "$prompt: " file_value

        if [ -z "$file_value" ]; then
            ((i--))
            echo "Entrada inválida. Le quedan $i intentos."
        elif [ -n "$extension" ] && [[ "$file_value" != *$extension ]]; then
            ((i--))
            echo "El fichero no tiene la extensión $extension. Le quedan $i intentos."
        elif [ "$check_exists" = "true" ] && [ ! -f "$file_value" ]; then
            ((i--))
        fi
    done
}
```

```

echo "☒ El fichero no existe. Le quedan $i intentos."
elif [ "$check_exists" = "false" ] && [ -f "$file_value" ]; then
    ((i--))
    echo "☒ El fichero ya existe. Le quedan $i intentos."
elif [ -n "$structure_regex" ] && grep -qvE "$structure_regex" "$file_value"; then
    ((i--))
    echo "☒ El contenido del fichero no sigue la estructura requerida. Le quedan $i intentos."
else
    eval "$file_var=\\"$file_value\\"
    return 0
fi

if [ "$i" -eq 0 ]; then
    return 1
fi
done
}

# Verifica la entrada que proporciona el usuario cuando se le pregunta (s/n)
# s -> 1, verdadero
# n -> 0, falso
validate_choice() {
    local prompt="$1"

    local i=3
    while [ "$i" -gt 0 ]; do
        read -rp "$prompt: " choice

        if [ "$choice" == "s" ]; then
            return 1
        elif [ "$choice" == "n" ]; then
            return 0
        else
            ((i--))
            echo "☒ Entrada inválida. Le quedan $i intentos."
        fi

        if [ "$i" -eq 0 ]; then
            return 2
        fi
    done
}

```

Al utilizar estas funciones que manejan la entrada por teclado que realiza el usuario el código de la opción 2 quedó mucho más legible, eficiente y reutilizable.

```

1) verify_input_file "emails_file" "Introduzca el nombre del fichero que contiene los correos electrónicos" "true" ".txt" "[0-9]"

if [ $? -eq 1 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú:" key
    clear
    continue
fi

verify_input_file "expressions_file" "Introduzca el nombre del fichero que contiene las expresiones sospechosas" "true" ".txt"

if [ $? -eq 1 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú:" key
    clear
    continue
fi

echo "¿Desea crear un fichero .freq donde se escribirá el análisis de los correos electrónicos?"
echo "Puede cargar este fichero mas adelante para calcular una predicción"
validate_choice "Introduzca una opción (s/n)"
create_freq_file=$?

if [ "$create_freq_file" -eq 1 ]; then
    verify_input_file "freq_file" "Introduzca el nombre del fichero (.freq) donde se escribirá el análisis de los correos electrónicos" "true" ".txt"

if [ $? -eq 1 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú:" key
    clear
    continue
fi

elif [ "$create_freq_file" -eq 2 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú:" key
    clear
    continue
fi

```

```

fi

total_emails=$(wc -l <"$emails_file")

# Analisis
# Se lee ID|Contenido del E-Mail|Resultado|Nada
i=0
while IFS="|" read -r email_id email_content spam_or_ham blank; do
    cleaned_email_content=$(clean_text "$email_content")
    total_expressions=$(echo "$cleaned_email_content" | wc -w | tr -d '[:space:]')
    freq_matrix["$i,0"]="$email_id"
    freq_matrix["$i,1"]="$total_expressions"
    freq_matrix["$i,2"]="$spam_or_ham"
    progressBar "Analizando E-Mail $email_id" "$i" "$total_emails"
    j=3
    while read -r expression; do
        cleaned_expression=$(clean_text "$expression")
        count=$(echo "$cleaned_email_content" | grep -wo "$cleaned_expression" | wc -l)
        freq_matrix["$i,$j"]="$count"
        ((j++))
    done <"$expressions_file"
    ((i++))
done <"$emails_file"

freq_matrix_rows=$i
cols=$j
analysis_completed=1

if [ "$create_freq_file" -eq 1 ]; then
    # Imprime la matriz freq en el fichero
    for ((i = 0; i < freq_matrix_rows; i++)); do
        for ((j = 0; j < cols; j++)); do
            echo -n "${freq_matrix["$i,$j"]}" >>"$freq_file"
        done
        echo >>"$freq_file"
    done
fi

# Recorre la matriz y cuenta los valores mayores que 0 en cada columna
for ((j = 3; j < cols; j++)); do
    column_counts[j]=0
    for ((i = 0; i < freq_matrix_rows; i++)); do
        if [[ "${freq_matrix["$i,$j]}" -gt 0 ]]; then
            ((column_counts[j]++))
        fi
    done
done
progressBar "[checkmark] Análisis completado con éxito" "$total_emails" "$total_emails"
echo
bannerColor "[document] Matriz de análisis guardada en: $(pwd)/$freq_file" "black" "*"
echo
read -rp "Introduzca cualquier tecla para regresar al menú:" key
;;

```

Para contar el numero total de palabras en cada correo electrónico despues de ser formateado utilicé el comando

`"$cleaned_email_content" | wc -w | tr -d '[:space:]'`, donde:

1. `echo "$cleaned_email_content"`: Imprime en la salida estándar el contenido de la variable `"$cleaned_email_content"`, es el "argumento" que se le pasará a la próxima tubería.
2. `|`: La tubería (pipe) se utiliza para redirigir la salida del comando anterior como entrada al siguiente comando. En este caso, la salida del comando "echo" se envía al siguiente comando.
3. `wc -w`: El comando "wc" (word count) se utiliza para contar palabras, líneas y caracteres en un texto. La opción "-w" le indica a "wc" que solo cuente las palabras en el texto.
4. `|`: Nuevamente, se utiliza una tubería para redirigir la salida del comando anterior al siguiente comando.
5. `tr -d '[:space:]'`: El comando "tr" se utiliza para realizar transformaciones en un texto. En este caso, se está utilizando para eliminar (borrar) todos los caracteres de espacio en blanco (espacios y tabulaciones) del texto. El argumento '`[:space:]`' especifica que se deben eliminar todos los caracteres de espacio en blanco.

Este ultimo comando no es realmente necesario, pero lo utilicé debido a qué en un cierto del desarrollo, al realizar `echo` de los algunos caracteres estos se imprian con espacios por delante y esto imposibilitaba las operaciones.

Para contar las apariciones de cada expresión en un correo electrónico utilicé el comando echo `echo`

```
"$cleaned_email_content" | grep -wo "$cleaned_expression" | wc -l
```

, donde:

1. `echo "$cleaned_email_content"`: Como en el comando anterior, este comando imprime en la salida estándar el contenido de la variable `"$cleaned_email_content"`.
2. `|`: La tubería se utiliza para redirigir la salida del comando anterior como entrada al siguiente comando.
3. `grep -wo "$cleaned_expression"`: El comando `grep` se utiliza para buscar patrones en un texto. Con las opciones `-wo`, `grep` busca palabras completas (`-w`) que coincidan exactamente con la expresión especificada `"$cleaned_expression"`. La opción `-o` indica que se deben mostrar solo las coincidencias, no las líneas completas.
4. `|`: Nuevamente, se utiliza una tubería para redirigir la salida del comando "grep" al siguiente comando.
5. `wc -l`: La opción `-l` le indica a `wc` que cuente las líneas en el texto. Dado que cada línea en la salida de `grep` corresponde a una coincidencia de la expresión, contar las líneas nos dará el número total de coincidencias.

Para obtener el total de correos electrónicos utilicé el comando `wc -l <"$emails_file"`, que proporciona el total de líneas del fichero ya que en cada línea se encuentra un correo electrónico. Solamente utilicé este dato para mandarselo como argumento a la función `progressBar()`, la cual imprime una barra de progreso.

2.3 Predicción

Esta quizás haya sido la opción que más he costado programar debido a lo que me costó comprender qué es lo que tenía que hacer y la complejidad que conlleva.

Al igual que en la opción anterior comencé programando una aproximación que presentaba mucho código redundante el cual logré optimizar (en menor medida) con las funciones mencionadas anteriormente y otra función que se encarga de verificar que el formato de un fichero (.freq o .tfidf) es correcto.

```
# Verifica si un fichero que contiene matrices, ya sea .freq o tfidf es correcto
validate_matrix_file() {
    local file="$1"
    local structure_regex="$2"
    while IFS= read -r linea; do
        # Verifica si la linea no está vacía
        if [[ -n $linea ]]; then
            # Verifica si el fichero tiene el formato correcto
            if [[ ! $linea =~ $structure_regex ]]; then
                return 0
            fi
        fi
    done <"$file"
    return 1
}
```

En primer lugar compruebo si el análisis se ha acabado de realizar y existe un análisis en la ejecución, en ese caso compruebo también si para ese análisis existe un fichero ".freq" para poder realizar la predicción, si no existe se le dará al usuario la opción de realizar la predicción cargando otro; en cambio, si existe; se le preguntará al usuario si desea utilizar dicho análisis o prefiere cargar un fichero de frecuencias ".freq".

```
2)
use_current_analysis=0
return_to_menu=0
freq_matrix_built=0
load_tfidf=0

if [ "$analysis_completed" -eq 1 ]; then

    if [ "$create_freq_file" -eq 0 ]; then
        echo "💡 Se ha encontrado un análisis en la ejecución pero no existe fichero de frecuencias (.freq)."
        echo "      Vuelva a realizar el análisis y asegúrese de crear el fichero de frecuencias (.freq)"
        validate_choice "      ¿Desea cargar un fichero de frecuencias para realizar una predicción? (s/n)"
        choice=$?

        if [ "$choice" -eq 1 ]; then
            use_current_analysis=0
        else
            read -rp "Introduzca cualquier tecla para regresar al menú:" key
            clear
            continue
        fi
    fi
fi
```

```

        elif [ "$create_freq_file" -eq 1 ]; then
            validate_choice "💡 Se ha encontrado un análisis en la ejecución. ¿Desea utilizarlo? (s/n)"
            use_current_analysis=$?
        fi

        if [ "$use_current_analysis" -eq 2 ]; then
            read -rp "Introduzca cualquier tecla para regresar al menú:" key
            clear
            continue
        fi
    fi

```

En el caso de que prefiera usar el análisis de la ejecución la matriz de frecuencias se copiará a la matriz tfidf, esta ultima vacía, para posteriormente realizar el cálculo del TF-IDF.

```

if [ "$use_current_analysis" -eq 1 ]; then
    #Variable auxiliar para contar el numero de filas que tendra la nueva matriz predicción
    k=0
    #Recorre la matriz de análisis y la copia a una nueva matriz para la predicción
    for ((i = 0; i < freq_matrix_rows; i++)); do
        email_id=${freq_matrix["$i,0"]}
        total_expressions=${freq_matrix["$i,1"]}

        if [[ total_expressions -gt 0 ]]; then
            for ((j = 0; j < cols; j++)); do
                tfidf_matrix["$k,$j"]=${freq_matrix["$i,$j"]}
            done
            ((k++))
        else
            echo "⚠️ Advertencia: El E-Mail $email_id está vacío, no se tendrá en cuenta en la calculación del TF-IDF."
        fi
    done
    tfidf_matrix_rows=$k
    freq_matrix_built=1
fi

```

Si por el contrario se desea cargar un fichero de frecuencias, el programa le pedirá al usuario dicho fichero y comprobará que es correcto. Además de verificar si existe un fichero ".tfidf" asociado a dicho fichero de frecuencias ".freq"

```

if [ "$analysis_completed" -eq 0 ] || [ "$use_current_analysis" -eq 0 ]; then
    return_to_menu=0
    file_ok=0
    tfidf_file=""
    i=3
    while [ "$i" -gt 0 ] && [ "$file_ok" -eq 0 ]; do
        read -rp "Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: " freq_file
        tfidf_file="${freq_file%.*}.tfidf"
        if [ -z "$freq_file" ]; then
            ((i--))
            echo "🔴 Entrada inválida. Le quedan $i intentos."
        elif [ ! -f "$freq_file" ]; then
            ((i--))
            echo "🔴 El fichero no existe. Le quedan $i intentos."
        elif [ -f "$tfidf_file" ] && [ -s "$tfidf_file" ]; then
            j=3
            input_valid=0

            validate_matrix_file "$tfidf_file" '^(-?[0-9]+(\.[0-9]*)(:-?[0-9]*(\..-[0-9]*))*):$'
            tfidf_file_is_valid=$?

            if [ "$tfidf_file_is_valid" -eq 0 ]; then
                echo "🔴 Advertencia: Se ha encontrado un fichero .tfidf no válido para el fichero .freq introducido."
                echo "Por favor, revise el fichero en el caso de que quiera utilizarlo para una nueva predicción."
                continue
            else
                echo "Se ha encontrado un fichero con extensión .tfidf para el fichero .freq introducido, ¿Desea cargarlo para"
                echo "s 🤝 El fichero con la matriz TF-IDF se utilizará para calcular una nueva predicción"
            fi
        fi
        if [ "$analysis_completed" -eq 0 ] || [ "$use_current_analysis" -eq 0 ]; then
            return_to_menu=0
            file_ok=0
            tfidf_file=""
            i=3
            while [ "$i" -gt 0 ] && [ "$file_ok" -eq 0 ]; do
                read -rp "Introduzca el nombre del fichero (.freq) del que quiere cargar el análisis: " freq_file
                tfidf_file="${freq_file%.*}.tfidf"
                if [ -z "$freq_file" ]; then

```

```

((i--))
echo "🔴 Entrada invalida. Le quedan $i intentos."
elif [ ! -f "$freq_file" ]; then
((i--))
echo "🔴 El fichero no existe. Le quedan $i intentos."
elif [ -f "$tfidf_file" ] && [ -s "$tfidf_file" ]; then
j=3
input_valid=0

validate_matrix_file "$tfidf_file" '^(-?[0-9]+(\.[0-9]*)(:-?[0-9]*)((\..-[0-9]*))*):$'
tfidf_file_is_valid=$?

if [ "$tfidf_file_is_valid" -eq 0 ]; then
echo "🔴 Advertencia: Se ha encontrado un fichero .tfidf no valido para el fichero .freq introducido."
echo "                Por favor, revise el fichero en el caso de que quiera utilizarlo para una nueva predicción"
continue
else
echo "Se ha encontrado un fichero con extensión .tfidf para el fichero .freq introducido, ¿Desea cargarlo para su uso?"
echo "s 🔴 El fichero con la matriz TF-IDF se utilizará para calcular una nueva predicción"
echo "n 🔴 El fichero con la matriz TF-IDF se borrará y se volverá a realizar el calculo del TF-IDF y la predicción"
fi

while [ "$j" -gt 0 ] && [ "$input_valid" -eq 0 ] && [ "$tfidf_file_is_valid" -eq 1 ]; do

read -rp "Seleccione una opción (s/n): " input

if [ -z "$input" ]; then
((j--))
echo "🔴 Entrada invalida. Le quedan $j intentos."
elif [ "$input" == "s" ]; then
load_tfidf=1
input_valid=1
file_ok=1
elif [ "$input" == "n" ]; then
rm "$tfidf_file"
load_tfidf=0
input_valid=1
file_ok=1
else
((j--))
echo "🔴 Entrada invalida. Le quedan $j intentos."
fi

done

elif [ -e "$freq_file" ]; then
# Compruebo si el fichero sigue el formato correcto
validate_matrix_file "$freq_file" '^(-?[0-9]+(:-?[0-9]+)*):$'
is_freq_file_valid=$?
if [ "$is_freq_file_valid" -eq 1 ]; then
file_ok=1
else
((i--))
echo "🔴 El fichero no sigue un formato valido. Le quedan $i intentos."
fi
fi

if [ "$i" -eq 0 ]; then
return_to_menu=1
break
fi

done

if [ "$return_to_menu" -eq 1 ]; then
read -rp "Introduzca cualquier tecla para regresar al menú:" key
clear
continue
fi

# Leer el archivo con la matriz .freq linea por linea
i=0
while IFS= read -r line; do
if [ -n "$line" ]; then
# Split de la linea en un array utilizando ":" como delimitador
IFS=":" read -ra elements <<<"$line"

email_id="${elements[0]}"
num_of_terms_in_email="${elements[1]}"

if [ "$num_of_terms_in_email" -gt 0 ]; then
for ((j = 0; j < ${#elements[@]}; j++)); do
tfidf_matrix["$i,$j"]="${elements[j]}"
done
fi
done
done

```

```

        ((i++))
    else
        echo "⚠️ Advertencia: El E-Mail $email_id está vacío, no se tendrá en cuenta en el cálculo del TF-IDF."
    fi
fi

done <"$freq_file"

tfidf_matrix_rows="$i"
cols=${#elements[@]}
freq_matrix_built=1
fi      echo "n 🚫 El fichero con la matriz TF-IDF se borrará y se volverá a realizar el calculo del TF-IDF y la predicción"
fi

while [ "$j" -gt 0 ] && [ "$input_valid" -eq 0 ] && [ "$tfidf_file_is_valid" -eq 1 ]; do

    read -rp "Seleccione una opción (s/n): " input

    if [ -z "$input" ]; then
        ((j--))
        echo "🔴 Entrada invalida. Le quedan $j intentos."
    elif [ "$input" == "s" ]; then
        load_tfidf=1
        input_valid=1
        file_ok=1
    elif [ "$input" == "n" ]; then
        rm "$tfidf_file"
        load_tfidf=0
        input_valid=1
        file_ok=1
    else
        ((j--))
        echo "🔴 Entrada invalida. Le quedan $j intentos."
    fi

    done

elif [ -e "$freq_file" ]; then
    # Compruebo si el fichero sigue el formato correcto
    validate_matrix_file "$freq_file" '^(-?[0-9]+(:-?[0-9]+)*):$'
    is_freq_file_valid=$?
    if [ "$is_freq_file_valid" -eq 1 ]; then
        file_ok=1
    else
        ((i--))
        echo "🔴 El fichero no sigue un formato valido. Le quedan $i intentos."
    fi
fi

if [ "$i" -eq 0 ]; then
    return_to_menu=1
    break
fi

done

if [ "$return_to_menu" -eq 1 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú:" key
    clear
    continue
fi

# Leer el archivo con la matriz .freq linea por linea
i=0
while IFS= read -r line; do
    if [ -n "$line" ]; then
        # Split de la linea en un array utilizando ":" como delimitador
        IFS=":" read -ra elements <<<"$line"

        email_id="${elements[0]}"
        num_of_terms_in_email="${elements[1]}"

        if [ "$num_of_terms_in_email" -gt 0 ]; then
            for ((j = 0; j < ${#elements[@]}; j++)); do
                tfidf_matrix["$i,$j"]="${elements[j]}"
            done
            ((i++))
        else
            echo "⚠️ Advertencia: El E-Mail $email_id está vacío, no se tendrá en cuenta en el cálculo del TF-IDF."
        fi
    fi

done <"$freq_file"

```

```

tfidf_matrix_rows="$i"
cols=${#elements[@]}
freq_matrix_built=1
fi

```

Una vez que se ha cargado ya sea el fichero ".freq" o el fichero ".tfidf" asociado se procedera bien a cargar el fichero ".freq" o el fichero ".tfidf" y a realizar el calculo del TF-IDF si es necesario, puesto que si se ha cargado el fichero TF-IDF y el usuario no desea volver a hacer el calculo solamente se lee del fichero y se copia a su matriz correspondiente.

Para realizar las operaciones he utilizado el comando `bc -l`, la cual se utiliza para cargar la biblioteca matemática de `bc`, que habilita funciones matemáticas adicionales, como el logaritmo natural y funciones trigonométricas.

```

if [ "$freq_matrix_built" -eq 1 ]; then
    # Crear el nuevo nombre de archivo con la extensión "tfidf"
    tfidf_file="${freq_file%.*}.tfidf"

    # Recorre la matriz y cuenta los valores mayores que 0 en cada columna
    for ((j = 3; j < cols; j++)); do
        column_counts[j]=0
        for ((i = 0; i < tfidf_matrix_rows; i++)); do
            if [[ "${tfidf_matrix[$i,$j]}" -gt 0 ]]; then
                ((column_counts[j]++))
            fi
        done
        done

    # Calcula TF-IDF
    echo "load_tfidf= $load_tfidf"
    echo "use_current_analysis= $use_current_analysis"
    if [ "$load_tfidf" -eq 0 ] || [ "$use_current_analysis" -eq 1 ]; then
        for ((i = 0; i < tfidf_matrix_rows; i++)); do
            email_id="${tfidf_matrix[$i,0]}"
            progressBar "Calculando TF-IDF para el E-Mail $email_id" "$i" "$tfidf_matrix_rows"
            for ((j = 3; j < cols; j++)); do
                occurrences=${tfidf_matrix[$i,$j]}
                total_terms_in_email=${tfidf_matrix[$i,1]}
                total_docs=$freq_matrix_rows
                docs_containing_term=${column_counts[j]}
                if [ "$docs_containing_term" -eq 0 ]; then
                    docs_containing_term=1
                fi
                tf=$(echo "scale=2; $occurrences / $total_terms_in_email" | bc)
                idf=$(echo "scale=2; l($total_docs/$docs_containing_term)/l(10)" | bc -l)
                tf_idf=$(echo "$tf * $idf" | bc -l)
                tfidf_matrix[$i,$j]="$tf_idf"
            done
            done

            progressBar "✅ TF-IDF calculado con éxito" "$tfidf_matrix_rows" "$tfidf_matrix_rows"
            echo
            echo
        # Si se ha cargado el TF-IDF solamente lo lee del fichero y lo copia
        # en el matriz tfidf
        elif [ "$load_tfidf" -eq 1 ]; then
            # Leer el archivo linea por linea
            tfidf_matrix_rows=$(wc -l <"$tfidf_file")"
            i=0
            while IFS= read -r line; do
                # Split de la linea en un array utilizando ":" como delimitador
                IFS=: read -ra elements <<<"$line"
                email_id=${elements[0]}"
                progressBar "Cargando TF-IDF para el E-Mail $email_id" "$i" "$tfidf_matrix_rows"
                for ((j = 0; j < ${#elements[@]}; j++)); do
                    tfidf_matrix[$i,$j]="${elements[j]}"
                done
                ((i++))
            done <"$tfidf_file"

            progressBar "✅ TF-IDF cargado con éxito" "$tfidf_matrix_rows" "$tfidf_matrix_rows"
            echo
            echo
        fi
    
```

Lo siguiente es realizar la predicción, por lo que se recorre la matriz linea por linea y se realiza una media de los valores TF-IDF calculados en cada celda, esta media se almacenará en una nueva columna (la última)

```

# Calcula la predicción. Resultado en nueva columna
for ((i = 0; i < tfidf_matrix_rows; i++)); do
    row_tf_idf_sum=0
    email_id="${tfidf_matrix[$i,0]}"
    progressBar "Calculando predicción para el E-Mail $email_id" "$i" "$tfidf_matrix_rows"
    for ((j = 3; j < cols; j++)); do
        tf_idf=${tfidf_matrix[$i,$j]}
        row_tf_idf_sum=$(echo "scale=2; $row_tf_idf_sum + $tf_idf" | bc)
    done
    average_tf_idf=$(echo "scale=2; $row_tf_idf_sum / (cols - 3)" | bc)
    if (($average_tf_idf > 0.3" | bc -l))); then
        tfidf_matrix[$i,$cols]=1
    else
        tfidf_matrix[$i,$cols]=0
    fi
done

progressBar "[checkmark] Predicción calculada con éxito" "$tfidf_matrix_rows" "$tfidf_matrix_rows"
echo

```

El ultimo paso es guardar la matriz en el fichero correspondiente

```

# Imprime la matriz final en el fichero .tfidf
for ((i = 0; i < tfidf_matrix_rows; i++)); do
    for ((j = 0; j < cols + 1; j++)); do
        echo -n "${tfidf_matrix[$i,$j]}:" >>"$tfidf_file"
    done
    echo >>"$tfidf_file"
done

bannerColor "[file] Matriz TF-IDF guardada en: $(pwd)/$tfidf_file" "black" "*"
echo
read -rp "Introduzca cualquier tecla para regresar al menú:" key
fi

```

2.4 Informes

En primer lugar, verifica si se ha realizado un análisis. De no ser así le avisará al usuario de que debe de relizar un análisis para poder visualizar los informes.

En la primera opción, se ejecuta un bucle que recorre el archivo que contiene las palabras. Dentro de este bucle, muestra el contenido de cada línea del archivo junto con una cantidad que se encuentra en la variable `${column_counts[i]}`, el cual contiene el numero de veces que aparece la expresión correspondiente en el documento.

En la segunda opción el usuario ingresa una expresión la cual se busca en el archivo que contiene las palabras. Si se encuentra la expresión, se busca su posición correspondiente en la matriz de frecuencias para verificar si existe algun correo con dicha expresión. En el caso de que exista se busca el correo por su ID en el fichero que contiene los correos y se muestran los 50 primeros caracteres.

En la tercera opción, busca un identificador de correo electrónico ingresado por el usuario en la matriz de frecuencias y si ese ID coincide con alguno cuenta recorriendo cuántas expresiones sospechosas aparecen en ese correo electrónico y muestra el resultado.

```

3)
if [ "$analysis_completed" -eq 0 ]; then
    echo "[error] Error: No se puede acceder a los informes sin realizar un análisis previo."
    read -rp "Introduzca cualquier tecla para regresar al menú:" key
    clear
    continue
else
    while true; do
        clear
        echo "Informes:"
        echo
        echo "1. Informe en formato fila/columna donde por cada término muestre en cuantos correos electrónicos del conjunto de"
        echo "2. Informe donde para un término particular, solicitado al usuario, se muestren los correos electrónicos donde ap"
        echo "3. Dado un identificador de correo electrónico, mostrar cuantos términos de los analizados aparecen."
        echo "4. Regresar al menu principal"
        echo
        read -rp "Seleccione una opción (1-4): " choice

        case "${choice}" in
            1)

```

```

i=3
while IFS= read -r expression; do
    echo "$expression ${column_counts[i]}"
    ((i++))
done <"$expressions_file"
read -rp "Introduzca cualquier tecla para regresar al menú de informes:" key
;;
2)
return_to_reports_menu=0
expression_found=0
expression_appears=0
i=3
while [ "$i" -gt 0 ] && [ "$expression_found" -eq 0 ]; do
    read -rp "Introduzca una expresión: " input
    if [ -z "$input" ]; then
        ((i--))
        echo "🔴 Entrada invalida. Le quedan $i intentos."
    else
        j=3
        while IFS= read -r expression && [ "$expression_found" -eq 0 ]; do
            if [ "$expression" == "$input" ]; then
                expression_found=1
            else
                ((j++))
            fi
        done <"$expressions_file"

        if [ "$expression_found" -eq 0 ]; then
            ((i--))
            echo "Expresión no encontrada, le quedan $i intentos."
        fi
    fi
    if [ "$i" -eq 0 ]; then
        return_to_reports_menu=1
        break
    fi
done

if [ "$return_to_reports_menu" -eq 1 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú de informes:" key
    clear
    continue
fi

# Recorrer la matriz de análisis para buscar el correo que contenga el término
for ((i = 0; i < freq_matrix_rows; i++)); do
    if [ "${freq_matrix["$i,$j"]}" -gt 0 ]; then
        expression_appears=1
        email_id="${freq_matrix["$i,0"]}"
        while IFS= read -r line; do
            # Verifica si la línea comienza con el id
            if [[ "$line" =~ ^$email_id ]]; then
                # Muestra los primeros 50 caracteres de la línea
                echo "${line:0:50}"
            fi
        done <"${emails_file}"
    fi
done

if [ "$expression_appears" -eq 0 ]; then
    echo "La expresión $input no aparece en ningún correo electrónico."
    echo
fi
read -rp "Introduzca cualquier tecla para regresar al menú de informes:" key
;;
3)
return_to_reports_menu=0
email_found=0
email_id=0
num_of_expressions=0
i=3
while [ "$i" -gt 0 ] && [ "$email_found" -eq 0 ]; do
    read -rp "Introduzca un identificador : " input
    if [ -z "$input" ]; then
        ((i--))
        echo "🔴 Entrada invalida. Le quedan $i intentos."
    elif [[ ! "$input" =~ ^[0-9]+$ ]]; then
        ((i--))
        echo "🔴 Entrada invalida. Le quedan $i intentos."
    else
        j=0
        while [ "$email_found" -eq 0 ] && [ "$j" -lt "$freq_matrix_rows" ]; do

```

```

email_id="${freq_matrix["$j,0"]}"
if [ "$email_id" -eq "$input" ]; then
    email_found=1
else
    ((j++))
fi
done

if [ "$email_found" -eq 0 ]; then
    ((i--))
    echo "Correo electrónico no encontrado, le quedan $i intentos."
fi

if [ "$i" -eq 0 ]; then
    return_to_reports_menu=1
    break
fi

done

if [ "$return_to_reports_menu" -eq 1 ]; then
    read -rp "Introduzca cualquier tecla para regresar al menú de informes:" key
    clear
    continue
fi

for ((i = 3; i < cols; i++)); do
    if [ "${freq_matrix["$j,$i"]}" -gt 0 ]; then
        ((num_of_expressions++))
    fi
done

echo "En el correo electrónico $email_id aparecen $num_of_expressions de un total de $((cols - 3)) expresiones sospechosas"
echo
read -rp "Introduzca cualquier tecla para regresar al menú de informes:" key
;;
4)

break
;;
*)
echo "default (none of above)"
;;
esac

done
fi
;;

```

4. Juego de pruebas

El juego de pruebas está definido junto con el manual de programador en el punto 3. Se pueden tomar como ficheros para los distintos casos de uso los ficheros presentados en ese apartado y comparar las salidas para verificar el correcto funcionamiento de la aplicación.

5. Bibliografia

Tf-idf

Tf-idf (del inglés Term frequency – Inverse document frequency), frecuencia de término – frecuencia inversa de documento (o sea, la frecuencia de ocurrencia del término en la colección de documentos), es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto. El valor tf-idf aumenta proporcionalmente al número de

w <https://es.wikipedia.org/wiki/Tf-idf>

TF IDF – Term Frequency – Inverse Document Frequency Text Classification by Dr. Mahesh Huddar

TF IDF – Term Frequency – Inverse Document Frequency Text Classification by Dr. Mahesh Huddar

This video discusses, how to extract the textual features that is tf-idf features from textual documents.

► <https://www.youtube.com/watch?v=OIV1Vblgz0I>

TF-IDF

Term Frequency – Inverse Document Frequency

Numerical Example

 <https://www.shellcheck.net/wiki/>

Esta ultima pagina web me ha ayudado mucho a aprender la sintaxis del lenguaje Bash.