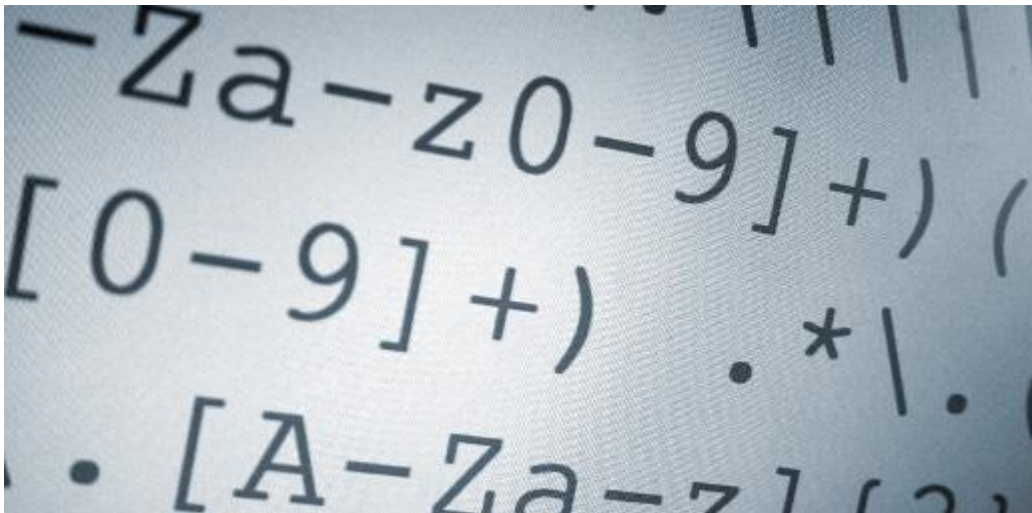


MÓDULO 3: OSINT

TAREA 1 - Expresiones regulares



Macià Salvà Salvà

ÍNDICE

ENUNCIADO	3
1. EXPLORACION DEL PDF	4
1.1 INVESTIGACIÓN	4
1.2 CONVERTIR PDF A TEXTO	5
2. HERRAMIENTA EGREP	6
2.1 PRIMER INTENTO	6
2.2 SEGUNDO INTENTO	8
2.3 TERCER INTENTO	9
3. CONCLUSIONES	10

ENUNCIADO

El alumno debería descargar el siguiente PDF del BOE y, adjuntar en el entregable del ejercicio, un listado con todos los DNI del documento, el número total de DNI encontrados, así como la expresión regular utilizada para extraer este tipo de información del documento PDF.

1. EXPLORACION DEL PDF

Para empezar he descargado el documento pdf del aula digital. Para asegurarme que es el fichero correcto he ejecutado el comando **hd** para verificar que se ha descargado en formato PDF.

```
hd BOE-A-2015-5834.pdf | head
```

```
(macia@kali)-[~/Downloads]
$ hd BOE-A-2015-5834.pdf | head
00000000  25 50 44 46 2d 31 2e 34 0a 25 e2 e3 cf d3 0a 32 |%PDF-1.4%. ....2|
00000010  31 20 30 20 6f 62 6a 0a 3c 3c 2f 50 61 72 65 6e |1 0 obj.<</Paren|
00000020  74 54 72 65 65 4e 65 78 74 4b 65 79 20 32 2f 52 |tTreeNextKey 2/R|
00000030  6f 6c 65 4d 61 70 20 33 20 30 20 52 2f 54 79 70 |oleMap 3 0 R/Typ|
00000040  65 2f 53 74 72 75 63 74 54 72 65 65 52 6f 6f 74 |e/StructTreeRoot|
00000050  2f 4b 5b 34 20 30 20 52 5d 2f 43 6c 61 73 73 4d |/K[4 0 R]/ClassM|
00000060  61 70 20 35 20 30 20 52 2f 50 61 72 65 6e 74 54 |ap 5 0 R/ParentT|
00000070  72 65 65 20 36 20 30 20 52 3e 3e 0a 65 6e 64 6f |ree 6 0 R>>.end|
00000080  62 6a 0a 38 20 30 20 6f 62 6a 0a 3c 3c 2f 54 79 |bj.8 0 obj.<</Ty|
00000090  70 65 2f 53 74 72 75 63 74 45 6c 65 6d 2f 4b 3c |pe/StructElem/K<
```

A continuación vamos hacer un estudio previo de la información que queremos extraer. Queremos extraer los documentos de identidad, esto incluye DNI y NIE. A su vez, si hay algún DNI que está mal formado, este también lo tendríamos que obtener.

1.1 INVESTIGACIÓN

En este apartado nos centramos en que queremos buscar. Concretamente el enunciado nos pide que busquemos DNI y NIE en todo el documento. Para ello vamos a ver en que formato están.

Suplente 3: AMARO MARTOS, ISMAEL. DNI 26048870J.
 Suplente 4: TODOROVA, POLINA PAULOVA. NIE X4652331Y.
 Suplente 5: VILLÉN MUÑOZ, IRENE MARÍA. DNI 77355980B.

En este ejemplo, observamos la presencia de DNI y NIE. El DNI consta de 8 números consecutivos seguidos por un carácter adicional. Por otro lado, el NIE está compuesto por 7 números consecutivos, finalizando con un carácter, y además comienza con otro carácter distinto.

Además hay un nombre donde se puede observar que el DNI es erróneo. Le falta el carácter final.

ALBALÁ SORIA, M.^a ÁFRICA. DNI 76923145.

1.2 CONVERTIR PDF A TEXTO

Para mejorar mi comodidad al trabajar, voy a emplear la herramienta llamada *AbiWord*. Esta herramienta tiene como objetivo convertir todo el contenido de un archivo *PDF* en texto plano.

```
abiword -t text BOE-A-2015-5834.pdf
```

```
(macia@kali)-[~/Downloads]
$ abiword -t text BOE-A-2015-5834.pdf

(macia@kali)-[~/Downloads]
$ ls
BOE-A-2015-5834.pdf  BOE-A-2015-5834.text  tor-browser  tor-browser-linux-x86_64-13.0.5.tar.xz  tor-browser-macos-13.0.5.dmg

(macia@kali)-[~/Downloads]
$
```

A continuación, podemos ver que se ha generado un archivo llamado **BOE-A-2015-5834.text**.

```
(macia@kali)-[~/Downloads]
$ head BOE-A-2015-5834.pdf
%PDF-1.4
%
21 0 obj
<</ParentTreeNextKey 2/RoleMap 3 0 R/Type/StructTreeRoot/K[4 0 R]
endobj
8 0 obj
<</Type/StructElem/K<</Pg 9 0 R/Obj 7 0 R/Type/OBJR>>/S/Form/P 10
endobj
11 0 obj
<</ParentTreeNextKey 2/RoleMap 3 0 R/Type/StructTreeRoot/K[8 0 R]

(macia@kali)-[~/Downloads]
$ head BOE-A-2015-5834.text
BOLETÍN OFICIAL DEL ESTADO
Núm. 126

Miércoles 27 de mayo de 2015

Sec. III. Pág. 45552

III. OTRAS DISPOSICIONES

MINISTERIO DE LA PRESIDENCIA

(macia@kali)-[~/Downloads]
$
```

Vamos a usar el comando **head** para comparar y comprender las diferencias entre distintos tipos de archivos. Cuando aplicamos este comando a un archivo en formato *PDF*, notamos que resulta ilegible debido a que contiene elementos específicos de formato propios de un *PDF*, como las cabeceras. A diferencia de un archivo de texto plano como el *.txt*, el *PDF* no se presenta de manera directa y sencilla para su lectura.

Sin embargo, al usar el comando **head** en el archivo que ha generado **aibword**, se puede notar que el texto contenido en su interior es legible.

2. HERRAMIENTA EGREP

Las expresiones regulares (regex) son patrones de búsqueda utilizados para identificar y manipular cadenas de texto. Son secuencias de caracteres que definen un conjunto de reglas de búsqueda para encontrar patrones dentro de texto. Permiten realizar búsquedas complejas, coincidencias o manipulaciones de cadenas de caracteres basadas en ciertos criterios, como secuencias específicas de caracteres, repeticiones, rangos, entre otros.

Egrep es una herramienta en sistemas basados en UNIX y Linux que se utiliza para realizar búsquedas de texto utilizando expresiones regulares extendidas. Es una variante de **grep**, que es una utilidad de línea de comandos utilizada para buscar texto dentro de archivos o flujos de texto. **Egrep** permite una mayor funcionalidad y versatilidad en la búsqueda al admitir expresiones regulares extendidas, lo que permite patrones más complejos y amplios en las búsquedas de texto.

2.1 PRIMER INTENTO

En el primer intento, he optado por el formato más básico y simple en la creación de una expresión regular. Me he enfocado en el DNI, que consta de 8 números. Por lo tanto, diseñé una expresión regular que identifica específicamente secuencias de 8 números consecutivos.

```
egrep "[0-9]{8}" BOE-A-2015-5834.text
```

```
(macia@kali)-[~/Downloads]
$ egrep "[0-9]{8}" BOE-A-2015-5834.text
Adjudicatario: PÉREZ ROIG, FRANCISCO JAVIER. DNI 44878140B.
Suplente 1: MATEOS LÓPEZ, ÓSCAR. DNI 47281731X.
Suplente 2: GONZÁLEZ ALEJO, AZOYE. DNI 78511780S.
Suplente 3: ROMERO VILCHEZ, BERTA. DNI 48959164F.
Suplente 4: RIPOLL ARCACIA, ELENA CRISTINA. DNI 48565665S.
Suplente 5: GARCÍA ALONSO, SERGIO. DNI 32067431A.
Adjudicatario: HERNÁNDEZ GÓMEZ DE CASO, MARÍA ISABEL. DNI 18451082E.
Suplente 1: CANO GARCÍA, SUSANA. DNI 26628204N.
Suplente 2: FERNÁNDEZ RODRÍGUEZ, ANA. DNI 28635052Y.
Adjudicatario: ALBALÁ SORIA, M.ª ÁFRICA. DNI 76923145.
Suplente 1: RIVEIRO RODRÍGUEZ, LORENA. DNI 78804483C.
Suplente 2: CASTRO DÍEZ, NATALIA. DNI 50229859K.
Suplente 3: SALAZAR PUERTA, SORAYA. DNI 49011750S.
Suplente 4: DE LA CRUZ GUTIÉRREZ, MARÍA. DNI 71228737Z.
Suplente 5: OYARBIDE MAGAÑA, ERNESTO EDUARDO. DNI 73509949V.
Suplente 3: TORREJÓN MORALES, SILVIA. DNI 02285005R.
Suplente 4: MARTÍN-ZARCO GALLEGU, ABDÓN. DNI 71226575Z.
Adjudicatario: VEGA GUERRERO, GRACIA MARÍA. DNI 02279784R.
Suplente 1: FRANGANILLO LOBATO, LAURA. DNI 72048845B.
Suplente 2: GARCÍA-MONTÓN GONZÁLEZ, PATRICIA. DNI 50898311R.
Suplente 3: AMARO MARTOS, ISMAEL. DNI 26048870J.
Suplente 5: VILLÉN MUÑOZ, IRENE MARÍA. DNI 77355980B.
Adjudicatario: GARCÍA BERNABÉ, ALBA. DNI 47097311G.
Suplente 1: ÁLVAREZ GUTIÉRREZ, ÁLVARO JESÚS. DNI 70257802R.
Suplente 2: MOYANO GARCÍA, LOURDES. DNI 30989411Q.
Suplente 3: MARTÍNEZ PÉREZ, MARÍA LLANOS. DNI 47081302A.
Suplente 4: BODOQUE FONT, ELENA. DNI 48603470P.
Suplente 5: ÁLVAREZ CALVO, M.ª VICTORIA. DNI 36135133V.
Adjudicatario: FOLGADO CARMONA, FERNANDO. DNI 28805154T.
Suplente 1: BENÍTEZ BODES, JUAN PABLO. DNI 80090151B.
Suplente 2: GARCÍA INFANTE, VANESSA. DNI 79343812E.
Suplente 3: LEIVA LÁZARO, FRANCISCO JAVIER. DNI 16605390Z.
Suplente 4: GÓMEZ TARANCÓN, ADRIÁN. DNI 72890230X.
Suplente 5: LÓPEZ PAZ, JESÚS. DNI 33540585F.
Suplente 6: MIGUÉLEZ MARTÍNEZ, DAVID. DNI 71014381H.
```

Si bien la situación actual no es completamente errónea, es evidente que el documento contiene tanto DNIs como NIEs. Sin embargo, estamos pasando por alto los NIEs en este análisis.

```
(macia@kali)-[~/Downloads]
$ egrep "[0-9]{8}" BOE-A-2015-5834.text | grep NIE

(macía@kali)-[~/Downloads]
$
```

Un ejemplo del NIE que tendríamos que ver sería este.

Suplente 5: PETROVICI, ZORANN. NIE X6989644J.

2.2 SEGUNDO INTENTO

En este segundo intento, hemos considerado también los NIEs. Para lograrlo, hemos creado una expresión regular que identifica números con una longitud de 6 u 8 dígitos seguidos de un carácter específico al final.

```
egrep "[0-9]{6,8}[A-Z]{1}" BOE-A-2015-5834.text
```

```
(macia@kali)-[~/Downloads]
$ egrep "[0-9]{6,8}[A-Z]{1}" BOE-A-2015-5834.text
Adjudicatario: PÉREZ ROIG, FRANCISCO JAVIER. DNI 44878140B.
Suplente 1: MATEOS LÓPEZ, ÓSCAR. DNI 47281731X.
Suplente 2: GONZÁLEZ ALEJO, AZOYE. DNI 78511780S.
Suplente 3: ROMERO VILCHEZ, BERTA. DNI 48959164F.
Suplente 4: RIPOLL ARCACIA, ELENA CRISTINA. DNI 48565665S.
Suplente 5: GARCÍA ALONSO, SERGIO. DNI 32067431A.
Adjudicataria: HERNÁNDEZ GÓMEZ DE CASO, MARÍA ISABEL. DNI 18451082E.
Suplente 1: CANO GARCÍA, SUSANA. DNI 26628204N.
Suplente 2: FERNÁNDEZ RODRÍGUEZ, ANA. DNI 28635052Y.
Suplente 1: RIVEIRO RODRÍGUEZ, LORENA. DNI 78804483C.
Suplente 2: CASTRO DíEZ, NATALIA. DNI 50229859K.
Suplente 3: SALAZAR PUERTA, SORAYA. DNI 49011750S.
Suplente 4: DE LA CRUZ GUTIÉRREZ, MARÍA. DNI 71228737Z.
Suplente 5: OYARBIDE MAGAÑA, ERNESTO EDUARDO. DNI 73509949V.
Suplente 3: TORREJÓN MORALES, SILVIA. DNI 02285005R.
Suplente 4: MARTÍN-ZARCO GALLEGU, ABDÓN. DNI 71226575Z.
Suplente 5: PETROVICI, ZORANN. NIE X6989644J.
Adjudicataria: VEGA GUERRERO, GRACIA MARÍA. DNI 02279784R.
Suplente 1: FRANGANILLO LOBATO, LAURA. DNI 72048845B.
Suplente 2: GARCÍA-MONTÓN GONZÁLEZ, PATRICIA. DNI 50898311R.
Suplente 3: AMARO MARTOS, ISMAEL. DNI 26048870J.
Suplente 4: TODOROVA, POLINA PAULOVA. NIE X4652331Y.
Suplente 5: VILLÉN MUÑOZ, IRENE MARÍA. DNI 77355980B.
Adjudicataria: GARCÍA BERNABÉ, ALBA. DNI 47097311G.
Suplente 1: ÁLVAREZ GUTIÉRREZ, ÁLVARO JESÚS. DNI 70257802R.
Suplente 2: MOYANO GARCÍA, LOURDES. DNI 30989411Q.
Suplente 3: MARTÍNEZ PÉREZ, MARÍA LLANOS. DNI 47081302A.
Suplente 4: BODOQUE FONT, ELENA. DNI 48603470P.
Suplente 5: ÁLVAREZ CALVO, M.ª VICTORIA. DNI 36135133V.
Adjudicatario: FOLGADO CARMONA, FERNANDO. DNI 28805154T.
Suplente 1: BENÍTEZ BODES, JUAN PABLO. DNI 80090151B.
Suplente 2: GARCÍA INFANTE, VANESSA. DNI 79343812E.
Suplente 3: LEIVA LÁZARO, FRANCISCO JAVIER. DNI 16605390Z.
Suplente 4: GÓMEZ TARANCÓN, ADRIÁN. DNI 72890230X.
Suplente 5: LÓPEZ PAZ, JESÚS. DNI 33540585F.
Suplente 6: MIGUÉLEZ MARTÍNEZ, DAVID. DNI 71014381H.

(macia@kali)-[~/Downloads]
$ egrep "[0-9]{6,8}[A-Z]{1}" BOE-A-2015-5834.text | grep NIE
Suplente 5: PETROVICI, ZORANN. NIE X6989644J.
Suplente 4: TODOROVA, POLINA PAULOVA. NIE X4652331Y.
```


El resultado es exitoso, aparentemente podríamos considerar esta práctica como finalizada y aceptada. Sin embargo, hemos identificado un error en un DNI específico:

ALBALÁ SORIA, M.^a ÁFRICA. DNI 76923145.

Se evidencia que le falta el último carácter que debería tener un DNI válido. Vamos a ajustar la expresión regular para que pueda aceptar esta entrada.

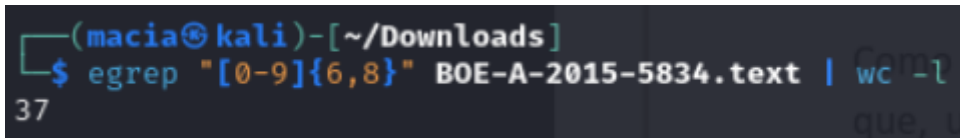
2.3 TERCER INTENTO

En este tercer intento, hemos logrado identificar los DNIs, NIEs y aquellos que contienen errores, como cuando falta la letra final. Para lograrlo, hemos conservado la expresión regular del intento anterior, simplemente eliminando el último carácter.

```
egrep "[0-9]{6,8}" BOE-A-2015-5834.text
```

```
(macia@kali)~[~/Downloads]
$ egrep "[0-9]{6,8}" BOE-A-2015-5834.text
Adjudicatario: PÉREZ ROIG, FRANCISCO JAVIER. DNI 44878140B.
Suplente 1: MATEOS LÓPEZ, ÓSCAR. DNI 47281731X.
Suplente 2: GONZÁLEZ ALEJO, AZOYE. DNI 78511780S.
Suplente 3: ROMERO VILCHEZ, BERTA. DNI 48959164F.
Suplente 4: RIPOLL ARCACIA, ELENA CRISTINA. DNI 48565665S.
Suplente 5: GARCÍA ALONSO, SERGIO. DNI 32067431A.
Adjudicataria: HERNÁNDEZ GÓMEZ DE CASO, MARÍA ISABEL. DNI 18451082E.
Suplente 1: CANO GARCÍA, SUSANA. DNI 26628204N.
Suplente 2: FERNÁNDEZ RODRÍGUEZ, ANA. DNI 28635052Y.
Adjudicataria: ALBALÁ SORIA, M.a ÁFRICA. DNI 76923145.
Suplente 1: RIVEIRO RODRÍGUEZ, LORENA. DNI 78804483C.
Suplente 2: CASTRO DÍEZ, NATALIA. DNI 50229859K.
Suplente 3: SALAZAR PUERTA, SORAYA. DNI 49011750S.
Suplente 4: DE LA CRUZ GUTIÉRREZ, MARÍA. DNI 71228737Z.
Suplente 5: OYARBIDE MAGAÑA, ERNESTO EDUARDO. DNI 73509949V.
Suplente 3: TORREJÓN MORALES, SILVIA. DNI 02285005R.
Suplente 4: MARTÍN-ZARCO GALLEGU, ABDÓN. DNI 71226575Z.
Suplente 5: PETROVICI, ZORANN. NIE X6989644J.
Adjudicataria: VEGA GUERRERO, GRACIA MARÍA. DNI 02279784R.
Suplente 1: FRANGANILLO LOBATO, LAURA. DNI 72048845B.
Suplente 2: GARCÍA-MONTÓN GONZÁLEZ, PATRICIA. DNI 50898311R.
Suplente 3: AMARO MARTOS, ISMAEL. DNI 26048870J.
Suplente 4: TODOROVA, POLINA PAULOVA. NIE X4652331Y.
Suplente 5: VILLÉN MUÑOZ, IRENE MARÍA. DNI 77355980B.
Adjudicataria: GARCÍA BERNABÉ, ALBA. DNI 47097311G.
Suplente 1: ÁLVAREZ GUTIÉRREZ, ÁLVARO JESÚS. DNI 70257802R.
Suplente 2: MOYANO GARCÍA, LOURDES. DNI 30989411Q.
Suplente 3: MARTÍNEZ PÉREZ, MARÍA LLANOS. DNI 47081302A.
Suplente 4: BODOQUE FONT, ELENA. DNI 48603470P.
Suplente 5: ÁLVAREZ CALVO, M.a VICTORIA. DNI 36135133V.
Adjudicatario: FOLGADO CARMONA, FERNANDO. DNI 28805154T.
Suplente 1: BENÍTEZ BODES, JUAN PABLO. DNI 80090151B.
Suplente 2: GARCÍA INFANTE, VANESSA. DNI 79343812E.
Suplente 3: LEIVA LÁZARO, FRANCISCO JAVIER. DNI 16605390Z.
Suplente 4: GÓMEZ TARANCÓN, ADRIÁN. DNI 72890230X.
Suplente 5: LÓPEZ PAZ, JESÚS. DNI 33540585F.
Suplente 6: MIGUÉLEZ MARTÍNEZ, DAVID. DNI 71014381H.
```

Como solución encontramos que hay 37 Documentos de identidad.

A terminal window with a dark background. The prompt is '(macia@kali)-[~/Downloads]'. The command entered is '\$ egrep "[0-9]{6,8}" BOE-A-2015-5834.text | wc -l'. The output is '37'.

```
(macia@kali)-[~/Downloads]  
$ egrep "[0-9]{6,8}" BOE-A-2015-5834.text | wc -l  
37
```

3. CONCLUSIONES

Utilizar expresiones regulares para identificar y extraer DNIs y NIEs de un documento de texto ofrece una solución eficiente. La flexibilidad y capacidad de especificar patrones específicos facilita la detección de estos números de identificación, ya que se pueden buscar coincidencias basadas en las características únicas de cada tipo de identificación.

Me ha llamado la atención retomar el uso de regex debido a mi experiencia previa en la universidad, donde las utilicé ampliamente durante mis asignaturas de compiladores. Me resulta fascinante emplear esta herramienta versátil, que puede ser empleada para desarrollar un lenguaje de programación, con el fin de identificar patrones específicos en documentos PDF, como la búsqueda de DNIs y NIEs.