La limpieza y transformación de datos con “expresiones regulares” o estandarizadas

Expositor: Jim Mallard
Secretary, International Association of Crime Analysts
Crime Analyst, Fort Collins (CO) Police Services
Jueves 28 de Julio, 2016

La limpieza y transformación de datos

con “expresiones regulares” o estandarizadas

Conceptos centrales

- ¿Cuáles son las “expresiones regulares” o estandarizadas?
- ¿Por qué debería utilizar “expresiones regulares”?
- ¿Cuándo debería utilizar “expresiones regulares”?
- Algunas reglas y sintaxis
- Ejemplos de cómo “expresiones regulares” pueden ayudar
- Fuentes y recursos

Objetivos

- Ayuda a entender el potencial de "expresiones regulares"
- Entender cómo éstas pueden ayudar a solucionar problemas de datos de manera más eficiente que cualquier otro método
- Objetivo no es hacer que usted se vuelva "experto" en "expresiones regulares"
  - **Difícil de aprender en un primer momento**
  - **Puede que no sea su trabajo**
- Tener conciencia de esto le ayudará a tener el apoyo de otros

Desafíos de limpieza de datos

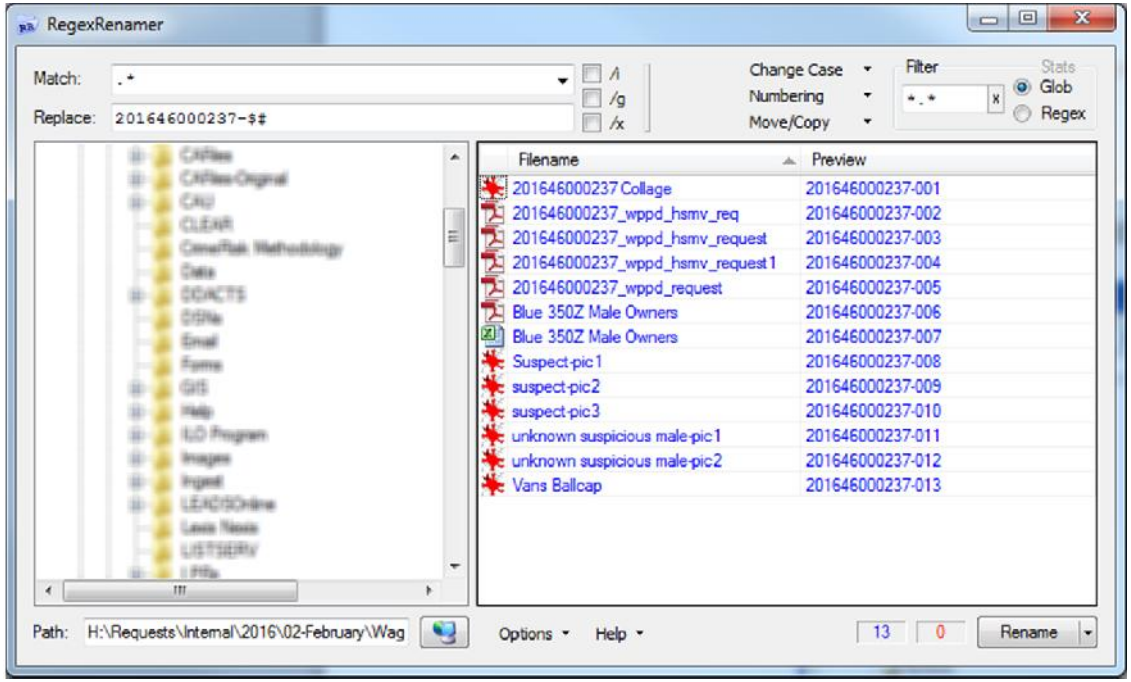
- ¿Cuáles son algunos problemas comunes que tiene con sus datos?
  - **Calles mal escritas?**
  - **Nombres mal escritos?**
  - **Números de teléfono que no están en el formato correcto?**
  - **…?**

¿Qué son las “expresiones regulares”?

- También llamadas **regex or regexp**
- Procesamiento de textos (notation)
  - **Disponible en la mayoría de los lenguajes y bases de datos**
  - **Java, Javascript, .NET, Perl, PHP, Python, Ruby, VBScript** y **Visual Basic**
- Bases de datos: **MySQL, Oracle**
- **Microsoft SQL Server** apoya alguna coincidencia de patrones utilizando la misma sintaxis, pero no todas las funciones de expresiones regulares
  - **Puede crear una «rutina» o CLR function (common language runtime) para exponer la funcionalidad de regex o expresiones regulares de .NET.**

¿Dónde puede usar “expresiones regulares” o regex?

- Como apoyo de muchos programas muy útiles
- Notepad ++
- **RegEx Renamer**
- Microsoft Access
- Microsoft Excel
- Ask/check



¿Qué son “regular expressions”?

- Como «**Buscar y reemplazar»**, pero mucho mejor
- **Describe** cadenas basadas en **patrones**
- Las palabras *no necesariamente* de forma literal
- Muy eficiente y flexible
- Una vez que aprenda las expresiones regulares, encontrará muchas oportunidades de usarlo en su trabajo cotidiano

¿Por qué usar “expresiones regulares”?

- Para limpiar, extraer, estandarizar datos «malos» o correctos para que puedan ser codificados geográficamente o analizados de manera más eficiente

  - **Nombres**
  - **Números de teléfono**
  - **Correos electrónicos**
  - **Los nombres de empresas**
  - **Cualquier texto puede ser limpiado**

Ejemplo en Facebook

- Facebook subpoena
  Peritaje de Facebook. Se refiere a un informe que elabora facebook respecto a todos los registros de un usuario que está siendo investigado por un delito.
- 5898 págs. de docs.
- Sólo se quería autor y el contenido del mensaje Fb

Facebook Business Record Page 2789

your fucking head off you actually ever fucking show up

Recipients Gerald Leroy Button Jr. (100001037200333)
William Cussins (100006384011614)
Author William Cussins (100006384011614)
Sent 2015-08-16 02:35:27 UTC
IP 2602:ae:10c4:6600::75
Deleted false
Body I got u

Recipients William Cussins (100006384011614)
Gerald Leroy Button Jr. (100001037200333)
Author Gerald Leroy Button Jr. (100001037200333)
Sent 2015-08-17 01:04:32 UTC
Deleted false
Body Hey what happened last night there buddy but you're going to fuck my life up there bro

Recipients Gerald Leroy Button Jr. (100001037200333)
William Cussins (100006384011614)
Author William Cussins (100006384011614)
Sent 2015-08-17 01:05:15 UTC
Deleted false
Body How to take care of Billy first even tried to pull a gun and knife but don't sweat it we will get it in I got nothing else to do

Recipients William Cussins (100006384011614)
Gerald Leroy Button Jr. (100001037200333)
Author Gerald Leroy Button Jr. (100001037200333)
Sent 2015-08-17 01:05:53 UTC
Deleted false
Body Whenever is clever

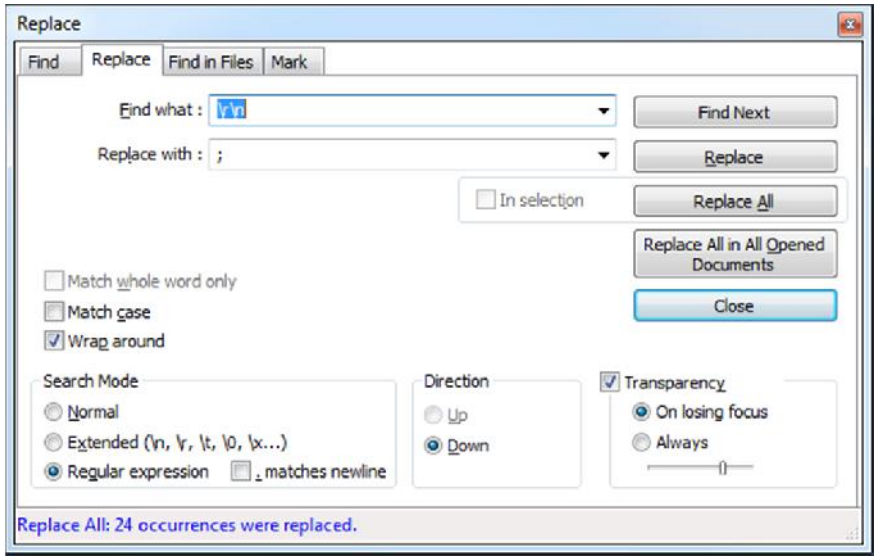
Ejemplo en Facebook

- Condensado a ~ 2300 páginas en un archivo de texto
  - **40% de la longitud del archivo original**
  - **Mucho más fácil de revisar**

```
1  Author Janette D Haidsiak
2
3  Wow. Really? Reread our convo. I dont know why you gotta call me that, but ok. Andsince when dont I bring money? Youve never fronted me shit
4
5  Author Janette D Haidsiak
6
7  Bitch and dumb fuck...that hurts
8
9  Author Janette D Haidsiak
10
11 Seriously Will. Wtf? Since when do you call me shit like that? After the shit I just toldyou was going on here, and uou come at me like that? I don't get it.
12
13 Author William Cussins
14
15 I'm under alot if stress
16
17 Author Janette D Haidsiak
18
19 Im sorry but I am too. And I didnt call you names. Ive been waiting for hours and stilldidnt call you names. That hurts
20
21 Author Janette D Haidsiak
22
23 If you cant bring me shit you could just say that. I asked you if you wanted me tomeet you?
```

Ejemplo direcciones de correos electrónicos

```
1  ApolloAbeytaLoera@cuvox.de
2  TicianoAbeytaValadez@rhyta.com
3  MahaAbregoOrnelas@cuvox.de
4  FabiolaAbreuBriones@fleckens.hu
5  FiorellaAbreuLerma@teleworm.us
6  FuensantaAbreuVega@fleckens.hu
7  AngeloAcevedoSoria@einrot.com
8  CalfucirAcevedoAdomo@jourrapide.com
9  LigioAcevedoTrevino@superrito.com
10 DioscoroAcostaAlvarez@einrot.com
11 MarjorieAcostaSantiago@dayrep.com
12 ShulamitAcostaVelasco@armyspy.com
13 ChloeAcunaMesa@armyspy.com
14 VertanAdomoGuerrero@superrito.com
15 KarimAgostoRangel@armyspy.com
16 NadinaAgostoAcuna@cuvox.de
17 PancraciaAgostoMenendez@einrot.com
18 LibiaAguayoHerrera@gustr.com
19 EdcoAguilarHerrera@cuvox.de
20 IoleAguilarJimenez@cuvox.de
21 MacraAguilarAguayo@fleckens.hu
22 ServandoAguilarLuevano@dayrep.com
23 AlaricoAguirreGracia@dayrep.com
24 LucindaAquirreAbrego@jourrapide.com
```




```
ApolloAbeytaLoera@cuvox.de;TicianoAbeytaValadez@rhyta.com;MahaAbregoOrnel
as@cuvox.de;FabiolaAbreuBriones@fleckens.hu;FiorellaAbreuLerma@teleworm.u
s;FuensantaAbreuVega@fleckens.hu;AngeloAcevedoSoria@einrot.com;CalfucirAc
evedoAdomo@jourrapide.com;LigioAcevedoTrevino@superrito.com;DioscoroAcost
aAlvarez@einrot.com;MarjorieAcostaSantiago@dayrep.com;ShulamitAcostaVelas
co@armyspy.com;ChloeAcunaMesa@armyspy.com;VertanAdomoGuerrero@superrito.c
om;KarimAgostoRangel@armyspy.com;NadinaAgostoAcuna@cuvox.de;PancraciaAgos
toMenendez@einrot.com;LibiaAguayoHerrera@gustr.com;EdcoAguilarHerrera@cuv
ox.de;IoleAguilarJimenez@cuvox.de;MacraAguilarAguayo@fleckens.hu;Servando
AguilarLuevano@dayrep.com;AlaricoAguirreGracia@dayrep.com;LucindaAguirreA
brego@jourrapide.com;
```

¿Por qué usar “regular expressions”?

- Transposición de nombres
  - **(\w+)\s(\w+)**
- Remoder caracteres especiales
  (,|/|-|#|@|""|\(|\))
- Encontrar / eliminar cualquier número de caracteres repetidos
  - **(#){2,}**

Antes	Despues
Apollo Abeyta	Abeyta, Apollo
Ticiano Abeyta	Abeyta, Ticiano
Maha Abrego	Abrego, Maha
Fabiola Abreu	Abreu, Fabiola
Fiorella Abreu	Abreu, Fiorella
Fuensanta Abreu	Abreu, Fuensanta
Angelo Acevedo	Acevedo, Angelo
Calfucir Acevedo	Acevedo, Calfucir
Ligio Acevedo	Acevedo, Ligio

Maestro Puig Valera #71	Maestro Puig Valera 71
Cruce Casa de Postas #62	Cruce Casa de Postas 62
La Fontanilla #8	La Fontanilla 8
Visitación de la Encina #67	Visitación de la Encina 67
Alvaro Cunqueiro #92	Alvaro Cunqueiro 92
C/ Amoladera #391	C/ Amoladera 391
Carretera Cádiz-Málaga #95	Carretera Cádiz-Málaga 95
La Fontanilla #69	La Fontanilla 69
Ctra de Siles #21	Ctra. de Siles 21

¿Por qué usar “regular expressions”?

- Eliminar espacios en blanco
  - Al inicio: **^\s+**
  - Al final Trailing: **\s+$**
  - En medio: **(\s){2,}**
- Nombre de lugar: LA Fitness
  - **^.*(L(\s|.)?A)(\s|.)(\s|-)?\s?FIT(NESS)?.*$**
- Matrículas parciales de vehículos
  - **C(\w+)7**
  - **43[53]8[ab]7**

¿Por qué usar “expresiones regulares”?

Dirty Location	Clean Location
J C PENNEY	JC PENNEY'S
J C PENNEY DEPARTMENT STORE	JC PENNEY'S
J C PENNEYS	JC PENNEY'S
J C PENNEY'S	JC PENNEY'S
J C PENNY	JC PENNEY'S
J C PENNYS	JC PENNEY'S
J.C. PENNEY	JC PENNEY'S
J.C. PENNEY #2685	JC PENNEY'S
J.C. PENNEY'S	JC PENNEY'S
JC PENNEY	JC PENNEY'S
JC PENNEY AT PARKS MALL	JC PENNEY'S
JC PENNEY AT THE PARKS MALL	JC PENNEY'S
JC PENNEY CLOTHING STORE	JC PENNEY'S
JC PENNEY DEPARTMENT STORE	JC PENNEY'S
JC PENNEY DEPT STORE	JC PENNEY'S
JC PENNEY PARKING LOT	JC PENNEY'S
JC PENNEY STORE	JC PENNEY'S
JC PENNEY, THE PARKS MALL	JC PENNEY'S
JC PENNEY,PARKS MALL	JC PENNEY'S
JC PENNEY,THE PARKS MALL	JC PENNEY'S
JC PENNEY/PARKS MALL	JC PENNEY'S
JC PENNEY/PARKS MALL-PKG LOT	JC PENNEY'S
JC PENNEY/THE PARKS MALL	JC PENNEY'S
JC PENNEY-PARKS AT ARLINGTON MALL	JC PENNEY'S
JC PENNEYS	JC PENNEY'S
JC PENNEY'S	JC PENNEY'S
JC PENNEYS PARKING LOT	JC PENNEY'S
JC PENNEY'S PARKS MALL	JC PENNEY'S
JC PENNEY'S, THE PARKS MALL	JC PENNEY'S
JC PENNEYS/THE PARKS MALL	JC PENNEY'S
JC PENNIES	JC PENNEY'S

Dirty Location	Clean Location
JC PENNY	JC PENNEY'S
JC PENNY AT PARKS MALL	JC PENNEY'S
JC PENNY DEPARTMENT STORE	JC PENNEY'S
JC PENNY STORE #2685-6	JC PENNEY'S
JC PENNYS	JC PENNEY'S
JC PENNY'S	JC PENNEY'S
JC PENNYS #2685	JC PENNEY'S
JC PENNYS AT THE PARKS MALL	JC PENNEY'S
JC PENNYS AT THE PARKS MALL IN ARLINGTON	JC PENNEY'S
JC PENNYS DEPARTMENT STORE	JC PENNEY'S
JCPENNEY	JC PENNEY'S
JCPENNEY/PARKS MALL	JC PENNEY'S
JCPENNEYS	JC PENNEY'S
JCPENNEY'S	JC PENNEY'S
JCPENNEYS DEPARTMENT STORE	JC PENNEY'S
JCPENNEY'S PARKS MALL	JC PENNEY'S
JCPENNYS	JC PENNEY'S
JCPENNY'S	JC PENNEY'S
PARKS MALL (JC PENNY)	JC PENNEY'S
PARKS MALL J C PENNEY PARKING LOT	JC PENNEY'S
PARKS MALL JC PENNEY PARKING GARAGE	JC PENNEY'S
PARKS MALL NEAR JCPENNEY PARKING LOT	JC PENNEY'S
PARKS MALL PARKING LOT NEAR JCPENNEY'S	JC PENNEY'S
PARKS MALL/J C PENNEY COVERED PARKING	JC PENNEY'S
PARKS MALL/JC PENNEY	JC PENNEY'S
PENNYS	PENNYS
THE PARKS MALL - J C PENNEY	JC PENNEY'S
THE PARKS MALL, BY JC PENNEY/MACY'S UPPER LEVEL PKG	JC PENNEY'S
THE PARKS MALL,2ND LEVEL NEAR JC PENNEY	JC PENNEY'S
THE PARKS MALL,IN FRONT OF JC PENNEY	JC PENNEY'S
THE PARKS MALL,NORTH SIDE OF JCPENNEY/MACY'S	JC PENNEY'S
THE PARKS MALL/JC PENNEYS	JC PENNEY'S

¿Por qué usar “expresiones regulares”?

^(N|S|E|W)(\s)(\d{1,4}) → $3$2$1

Dirty Address	Clean Address
W 2246 KENTUCKY AV	2246 W KENTUCKY AVENUE
W 225 CANTON AV	225 W CANTON AVENUE
W 225 FAIRBANKS AV	225 W FAIRBANKS AVENUE
W 225 LYMAN AV	225 W LYMAN AVENUE
W 2250 AFIRBANKS AV	2250 W AFIRBANKS AVENUE
W 2250 FAIRBANKS AV	2250 W FAIRBANKS AVENUE
W 2250 FAIRBANKS AV APT 1	2250 W FAIRBANKS AVENUE
W 2268 FAIRBANKS AV	2268 W FAIRBANKS AVENUE
W 227 CANTON AV	227 W CANTON AVENUE
W 227 NEW ENGLAND AV	227 W NEW ENGLAND AVENUE
W 227 NEW ENGLAND AV APT B	227 W NEW ENGLAND AVENUE
W 227 NEW ENGLAND AV APT B&C	227 W NEW ENGLAND AVENUE
W 227 PARK AV	227 W PARK AVENUE
W 228 WELBOURNE AV	228 W WELBOURNE AVENUE
W 2286 FAIRBANKS AV	2286 W FAIRBANKS AVENUE
W 229 FAIRBANKS AV	229 W FAIRBANKS AVENUE
W 230 CANTON AV	230 W CANTON AVENUE
W 230 READING WY	230 W READING WY
W 2300 FAIRBANKS AV	2300 W FAIRBANKS AVENUE
W 2300 FAIRBANKS AV APT BLK	2300 W FAIRBANKS AVENUE
W 231 KINGS WAY	231 W KINGS WAY
W 231 KINGS WY	231 W KINGS WY
W 231 PARK AV	231 W PARK AVENUE
W 231 ROCKWOOD WY	231 W ROCKWOOD WY
W 2324 FAIRBANKS AV	2324 W FAIRBANKS AVENUE

Dirty Phone	Clean Phone
201-243-8949	(201) 243-8949
2019279797	(201) 927-9797
202-276-9064	(202) 276-9064
202-732-3193	(202) 732-3193
202-842-9330	(202) 842-9330
205-803-8858	(205) 803-8858
206-233-5027	(206) 233-5027
206-262-2457	(206) 262-2457
206-389-3827	(206) 389-3827
206-433-1822	(206) 433-1822
206-550-6412	(206) 550-6412
207-874-8553	(207) 874-8553
2082014648	(208) 201-4648
208-570-6061	(208) 570-6061
21- 221667	
210-207-7680	(210) 207-7680
210-247-7649	(210) 247-7649

¿Por qué usar “expresiones regulares”?

Call Summary	Extracted Phone	Extracted Phone (Formatted)
** FRAUDULENT USE OF IDENTITY ** SOMEONE HAS USED COMPS BAMK CARD NUMBER TO MAKE A PURCHASE, OL \NAME:CARL HOLLINGS \PH:817 681 7160	817 681 7160	(817) 681-7160
UUMVVEH TAKEN TODAY FROM LOC BY COUSIN....18 YOA AND HAS NOT RETURNED VEH. VEH IS A CHAMPAGNE, 88, CHEVY PRISM, LP #231SFT. OL. \NAME:GUERRANT, ANTOINE \PH:817-468-0116	817-468-0116	(817) 468-0116
15368 /02-20-07/ACTIVE,ARNOLD LOGISTICS,717-730-5212,...AUD, DOORS 10-14, NO ANS, PS#817-652-1295, C/S. OL \NAME:ADT OPR#TIARA \PH:877-238-7739	717-730-5212	(717) 730-5212
22679 /09-27-07/ACTIVE,ROSALINDA B LARA,(817)784-3726 ,CELL PH#817-999-4995 SPOUSE BUS PH#972-243-0977, AUD, FRONT AND GARAGE DOORS, FEMALE-GLORIA DOES NOT HAVE PROPER, WILL ATC REP. CS OL \NAME:HOME	(817)784-3726	(817) 784-3726
33412 /05-22-07/ACTIVE,HECTOR L MATOS,817-557-4089,WWW MONITRONICS.COM--E MAIL--SUPPORT@MONITRONICS.COM AUD, GARAE DOOR, NO ANS, C/S OL \NAME:MONITRONICS - CHRIS \PH:800-852 2511	817-557-4089	(817) 557-4089
51989 /12-12-07,SHANDRA DALE SMIITH,(817)417-8243 AUD, FRONT DOOR, ANS MACH, C/S OL \NAME:ADT - GLY \PH:877 238 7730	(817)417-8243	(817) 417-8243
AUD, GARAGE DOOR (POSSIBLY THE DOOR BETWEEN THE GARAGE AND THE HOUSE) AT THELISA TAYLOR RESD, #817-419-9997. A MALE NAMED BRANDON TAYLOR IS INSIDE W/THEINCORRECT CODE. THERE IS NO PERMIT ON FILE. REF	817-419-9997	(817) 419-9997
CAR KEYS LOST WHILE AT WORK AT LOC ADDRESS. OL \PH:817/791 4084 \SOURCE:RESD	817/791 4084	(817) 791-4084
COMP ADV WIFE REFUSING TO LEAVE COMP IN. BOTH PARTIES LIVE TOGETHER. NO WEAPS. \SOURCE:WRLS		
COMP JUST CAME HOME & FOUND MOTHER DEAD \PH:817.673.6273 (MTF)	817.673.6273	(817) 673-6273
COMPS BOYFRIEND ASSAULTED COMP LAST NIGHT.. NO EMS NEEDED.. SUSP VEH RED 92 CADILLAC ELDORADO 2 DR UNK LP#.. SUSP IS B/M, 42YOA, 6'3/215, SHORT BLK HAIR, GOATEE, JEFFREY MORGAN DOB 7/20/64. \NAME:TAM		
COMPS GIRLFRIEND IS MISSING, LS 122306, VIRGINA MOSS W/F 072264, 411/105, BROWNISH RED HAIR, LIVES AT 2900 FOREST HOLLOW LN #2415 SUSP FLEW TO CHARLESTON SC & DID NOT RETURN, DUE BACK LAST NIGHT, VOI		
COMPS MOTHER TOOK OD OFF PILLS \NAME:CURTIS BURTON \SOURCE:RESD (MTF)		
COMPS PURSE STOLEN FROM COMPS VEH AT APPROX 1700. O/L \NAME:DOROTHY JACKSON \PH:817-801-6966 \SOURCE:VOIP	817-801-6966	(817) 801-6966
FORCED ENTRY TO STORAGE UNIT, OCCURRED BETWEEN NOVEMBER AND DECEMBER, OLD CHECKBOOK WAS STOLEN AND CHECKS ARE BEING USED, COMP WILL BE IN GRAY GEO STORM PARKED BY THE FRONT OFFICE IN APPROX 5-10 MIN,		
HANGUP. NO VOICE CONTACT. \SOURCE:RESD		
JASON HOLLAND W/M 33YOA DOB 062473, 508/170, BRO HAIR, BLU EYES, UNK CLOTHING. COMP ADV SUSP IS BREAKING GLASS, CUSSING, KICKING AND THROWING THINGS \NAME:DIANA SEYMOUR \PH:2147276881 (MTF)	2147276881	(214) 727-6881
MOTHER HIT COMP ON BACK OF HEAD WHEN SHE WAS TRYING TO LEAVE. NO EMS NEEDED \NAME:WILLIAMS,SHILANTA \SOURCE:WRLS		
NO FURTHER INFORMATION		

Sintaxis

Anchors (*Anclajes*)

- Coincidencia de *positions* dentro de una cadena, no caracteres
- ^ coincidencia al inicio de una línea
  - **^antes** matches "Antes que te cases, mira lo que haces" pero no "mira lo que haces antes que te cases"
- $ coincidencia al final de una línea
  - **haces$** matches "Antes que te cases, mira lo que haces" but not "mira lo que haces antes que te cases"
- Nota: Ambos coincidirían con palabras parciales a no ser que se incluya \ b (**^antes\b** or **\bhaces$**)

Adaptación o coincidencia del elemento precedente

- * Coincidencia del elemento precedente “cero” o más veces
- ? Coincidencia del elemento precedente “cero” o una vez
  - **Igual que "opcional“**
- + Coincidencia de elementos anteriores una o más veces

Adaptación o coincidencia del elemento precedente

- () crear un subgrupo dentro del patrón
- {x} coincidencias de elementos precedentes exactamente por X veces
- {x,y} coincidencias de elementos precedentes entre x e y veces
- {x,} coincidencias de elementos precedentes al menos x veces
- {,y} coincidencias de elementos precedentes entre “cero” e **Y** veces

Sintaxis

Set de Caracteres

- El período o “punto” (.) coincide con cualquier caracter excepto un caracter de nueva línea
  - **Así. * Coincidirá con cualquier número de caracteres**

- **A|B coincidencias "A" o "B"**

Tipos de Abreviaturas de Caracteres

- \w y \W
  - **\w = “palabras” caracteres (letras, dígitos, el subrayado)**
  - **\W = nada que no sea un caracter de palabra**
- \d and \D
  - **\d = digito**
  - **\D = nada que no sea un dígito**
- \s and \S
  - **\s = espacio en blanco (espacios, tabulaciones, salto de línea)**
  - **\S = nada que no sea un espacio en blanco**

Sintaxis

Set de Caracteres

- [a-n], [0-5] coincide con cualquier caracter único dentro del rango de letras o números entre paréntesis
- [^a-n], [^0-5] Coincide con cualquier caracter individual que no está dentro del rango de letras o números entre paréntesis

Sintaxis

Set de Caracteres

- abc, ABC coincidencia "abc" o "ABC"
- [abc] coincidencia con cualquier carácter único dentro de los paréntesis“a” or “b” or “c”
- [^abc] coincidencia con cualquier carácter único que no este dentro de los paréntesis “d” hasta “z”

Metacaracteres

12 caracteres con significados especiales

- El signo de intercalación: ^
  - **Significa "en el inicio de la cadena"**
  - **También niega lo que sigue, si está entre paréntesis**
- El signo del dolar: $
  - **Significa "al final de la cadena"**
- El período o un punto: .
  - **Coincide con cualquier carácter excepto una nueva línea de caracteres**

Metacaracteres

12 caracteres con significados especiales

- La barra vertical de símbolo de tubo: |
  - **Separa los elementos de una lista**
  - **Lo mismo que decir "o"**
- El signo de interrogación: ?
  - **Coincidencia de elementos precedentes "cero" o una vez**
  - **Lo mismo que decir que un elemento es "opcional"**
- El asterisco o inicio: *
  - **Coincidencia de elementos precedentes "cero" o más veces**
- El signo más: +
  - **Coincidencia de elementos precedentes "una" o más veces**

Metacaracteres

12 caracteres con significados especiales

- El paténtesis de inicio: (
  - **Inicio de un grupo**
  - **(a|b|c), (lunes|martes|miércoles)**
- El parentesis de cierre o final:)
  - **El término de un grupo**
- El inicio de un corchete: [
  - **Indica el inicio de un conjunto de caracteres**
  - **[abc]**

Metacaracteres

12 caracteres con significados especiales

- La llave de apertura: {
  - **Indica el inicio de un rango**
  - **{2,5}**
- La barra invertida: \
  - **Por sí mismo, que se utiliza para "evitar" otros caracteres especiales**

Ejemplo de nombre

- Patrón: **^arell[aeu]+no$**
- Coincidencia
  - **Arellano**
  - **Arelleno**
  - **Arelluno**
- No coincide
  - **Arellino**

Ejemplo de nombre

- Patrón: (po)?zuel(o|a)\sdel\srey
- Coincidencia
  - **Pozuelo del Rey**
  - **Zuelo del Rey**
  - **Pozuela del Rey**
- No coincide
  - **Pozuelo de Rey**
    - Pero se puede modificar: (po)?zuel(o|a)\sde(l)?\srey

Preguntas que Ud. tienen que hacerse cuando utiliza “expresiones regulares”

- ¿Será esta la coincidencia que yo quiero?
- ¿Coincide esto razonablemente con cualquier cosa que no quiero?
  - **Patrón puede resultar en falsos positivos pero no hay problema si la coincidencia no deseado es improbable que exista en los datos**
- ¿Cuáles son mis supuestos acerca del patrón y son razonables?
  - **Suposición de que los 5 primeros caracteres de un código postal de Estados Unidos son dígitos (*7 en Chile*)**
  - **El inicio de una dirección en Estados Unidos contiene dígitos**

“Expresiones regulares”

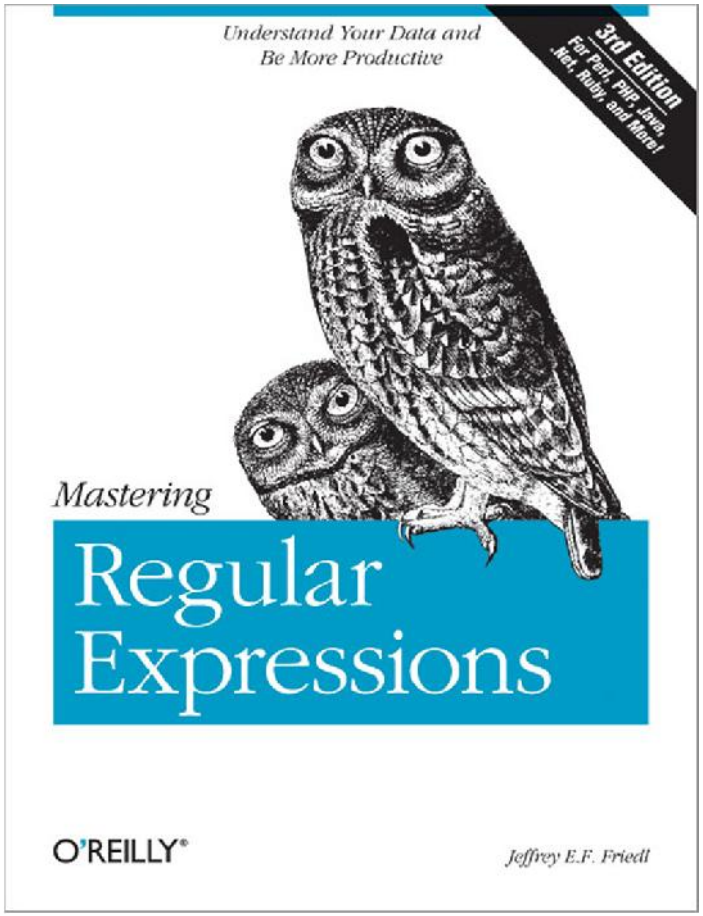
- En parte arte, en parte ciencia
  - **Al igual que escribir**
  - **Algunos “hermosos”, algunos difíciles de leer**
- No hay tal cosa como una expresión "perfecta"
- Algunos son mejores que otros
- El objetivo o meta es reducir los “falsos positivos”

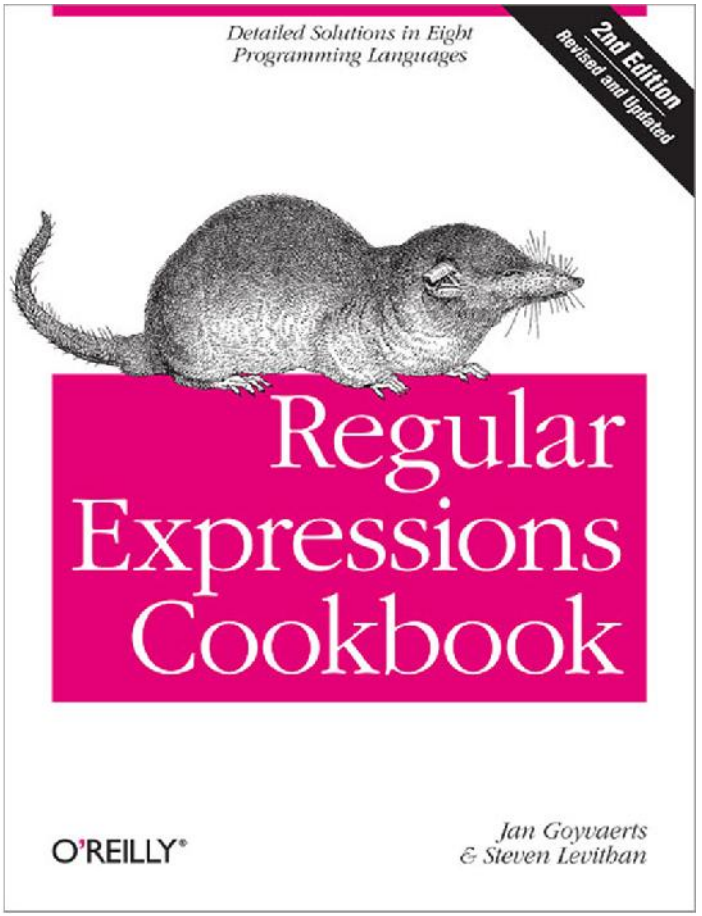
Recursos y fuentes

VBA Funciones

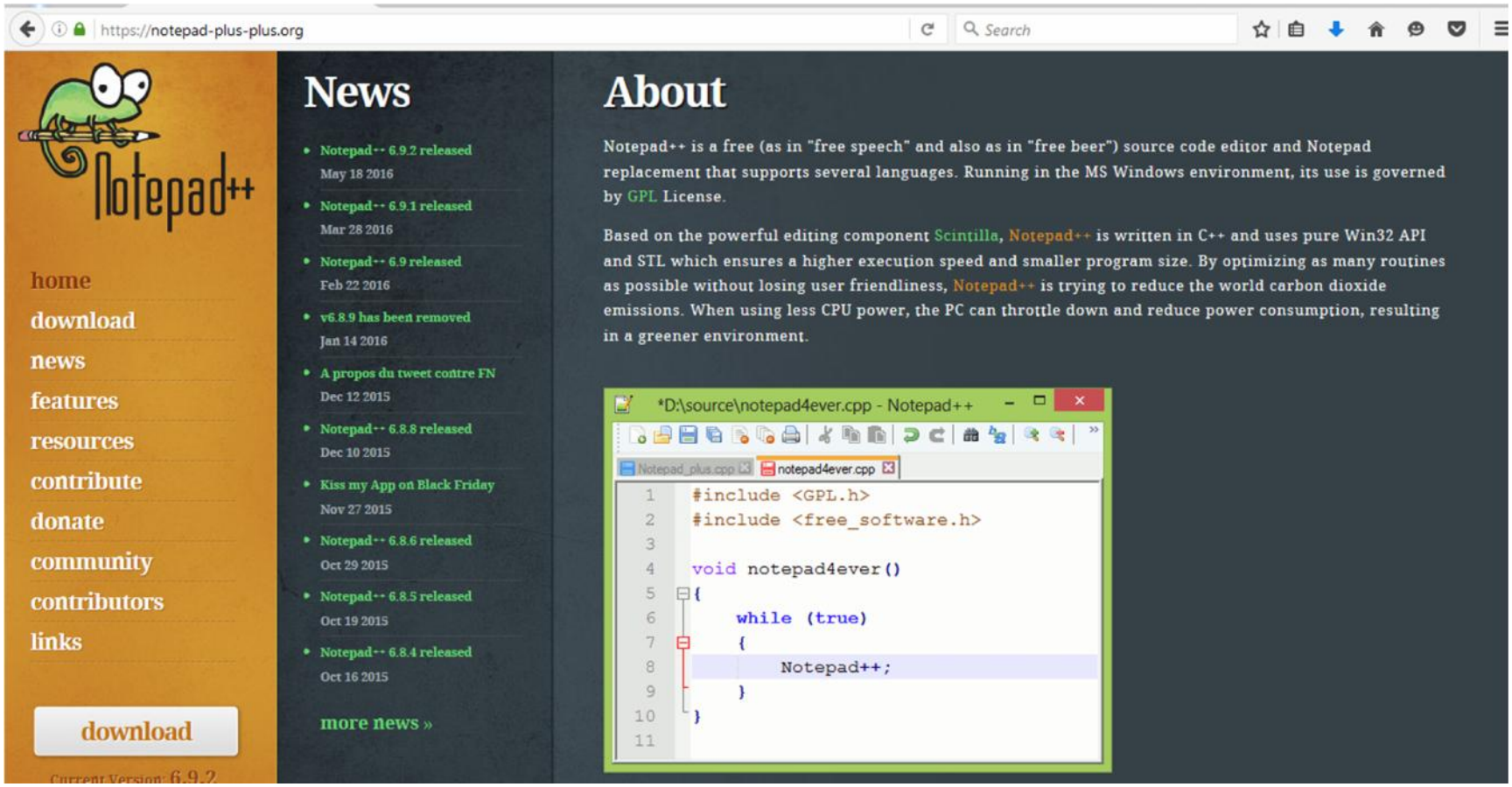
- http://tinyurl.com/regex-for-access
- http://tinyurl.com/regex-for-excel
- **Copiar el código**
- **Pegar en módulos VBA y ahorra**
- **Utilice las funciones de Access y Excel!**
- Lea comentarios del código de instrucciones y consejos

Fuentes

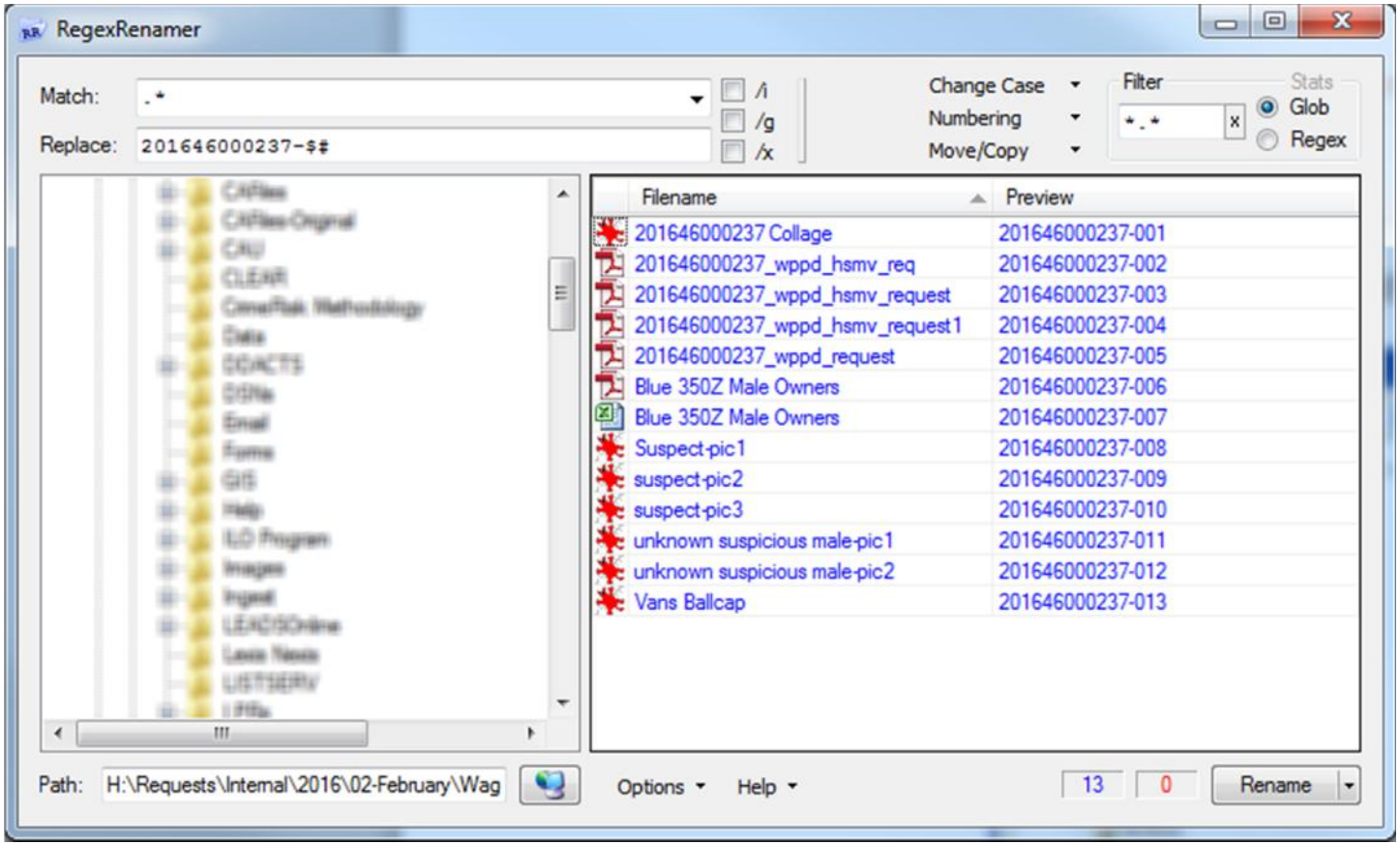




Notepad Plus Plus



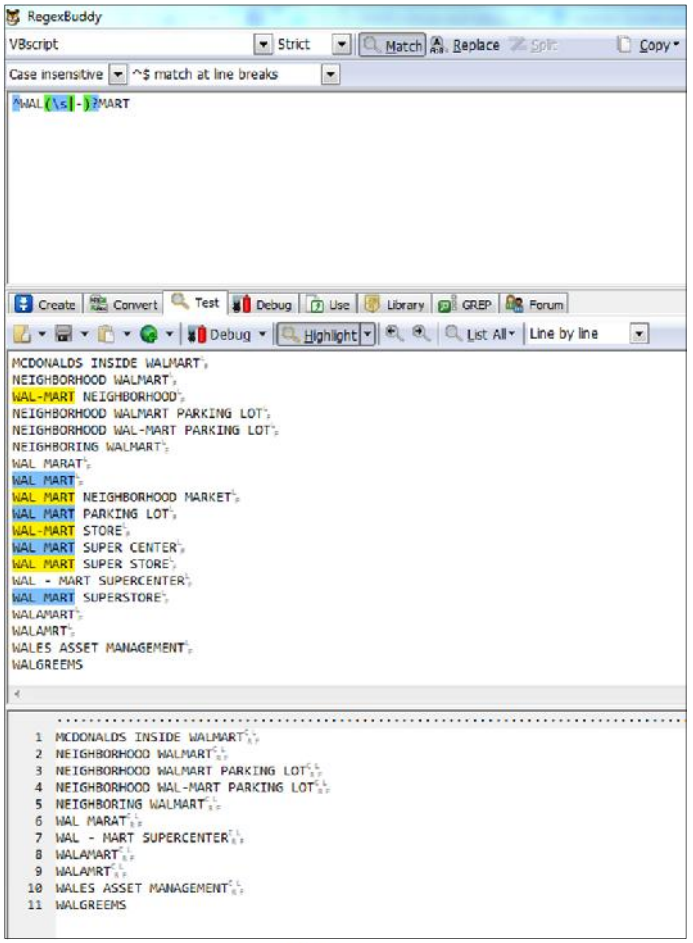
RegEx Renamer



RegEx Buddy

www.regexbuddy.com

- Se utiliza para crear y probar expresiones regulares en los datos de una muestra



Gracias!

Jim Mallard

Vice President of Administration
International Association of Crime Analysts

9218 Metcalf Ave #364, Overland Park, KS 66212

1.800.609.3419

iaca@iaca.net

http://www.iaca.net

@crimeanalysts

International Association of Crime Analysts

International Association of CRIME ANALYSTS