Tras un tiempo sin publicar ninguna nueva entrada, en esta ocasión, voy a compartir una**Chuleta / Guía / Tutorial de Open Refine (Google Refine)**.
No se trata de un completo tutorial de esta fabulosa herramienta sino que más bien se trata de una 'hoja de trabajo' que con frecuencia tengo 'a mano' durante la mayoría de primeras acciones en los proyectos de SEO, SEM y Analítica Web que realizo.

Chuleta / Guía / Tutorial de Open Refine (Google Refine)

OpenRefine (ex-Google Refine) es una potente herramienta multiplataforma para trabajar con datos desordenados, permitiendo limpiarlos y transformarlos. No es un servicio web sino una aplicación de escritorio, por lo que nuestros datos están seguros, aunque interactúe vía web (la aplicación se abre en el navegador).
Desde el 2 de Octubre 2012, Google no está apoyando activamente este proyecto que ha sido renombrado a **OpenRefine**. El desarrollo del proyecto, la documentación y la promoción está ya plenamente apoyados por voluntarios.

Generalidades

- Fácilmente se puede descargar e instalar desde su web: http://openrefine.org/
- Podemos concatenar varias sentencias, por ejem, *value.replace ("total", "").replace ("-", "")*.
- Podemos tener varias 'facet' al mismo tiempo obteniendo subconjuntos de subconjuntos de registros.
- Si nos equivocamos: 'retoceder' (Undo/Redo).
- En las expresiones podemos usar comillas dobles o simples (simples para texto).
- Utiliza lenguaje GREL en el que podemos usar expresiones regulares, concretamente en sus funciones: replace, match, partition, rpartition y split.

Filtrar: explorar, ver, encontrar

Se trata de aplicar uno o más de los siguientes filtros. Clic en 'count' para ver nº de repeticiones.
Filtrar: *Text filter* y empezar a escribir los valores que queremos encontrar.
Para localizar determinados valores: *facet>custom text facet*

- Ejem: escribir *value.contains("million")* para localizar estos valores
- Ejem: escribir *toNumber(value.replace(" million", ""))*1000000* para eliminar 'million' y multiplicar por 1000 el contenido restante.

Ver valores duplicados: *Facet>Customized facets>Duplicates facet*

Ver las partes (que se repiten) de los valores de una columna: *Facet>Customized facets>Word facet*

Ver todos los valores de una columna: *Facet>Text facet*

Ver (y cambiar o agrupar) variantes similares de valores: *edit cells>cluster and edit* (ver 'transformar datos').

Obtener celdas vacías de una columna (para por ejem eliminar esos registros) :*Facet>Customized facets>Facet by blank* y clic en 'true'

- Clic en la columna principal y clic en ordenar (*Sort>Sort...*)
- Clic en *Sort>Reorder rows permanently* situado en el nuevo menú superior que aparece
- Situar las celdas vacías al final: *Edit cells>Blank down*
- Filtramos para sólo mostrar los valores vacíos: *Facet>Customized facets>Facet by blank* y clic en '*true*'
- Para eliminar (ver 'transformar datos)

Explorar datos con gráficos de dispersión

En la columna que nos interese: *Facet>Scatterplot facet* (clic en '*log*' para verlo mejor). Ello muestra las relaciones entre todos los valores numéricos en cada una de las columnas.

Clic en uno de los gráficos para filtrar/mostrar esos registros.

Clic-arrastrar sobre el gráfico de la izq para seleccionar un grupo de registros y mostrarlos.

Transformar datos

Tras aplicar 1 ó más filtros:

- en el panel izquierdo podemos editar lo que interese. Por ejem: cambiar conjunto de valores tras filtrar por nombre (renombrar, renombrar uno para que se fusione con otro (agrupar), etc).
- también podemos hacerlo en cualquier celda de la dcha (datos filtrados) y aplicar los cambios a todas las instancias o celdas similares.

Transformaciones con*Edit cells>transform*. Escribir:

- *value.unescape('url')* para eliminar los caracteres raros de una url
- *value.replace("+","")* para eliminar el signo '+' **(reemplazar valores)**
- *replace(value, "+", "")* para buscar, entre los valores, el signo '+' y reemplazarlo por cadena vacía **(reemplazar valores)**
- *'http://'+ cells['nombre_columna'].value +'.com'* para obtener 'http://valor.com'**(añadir caracteres)**
- *'Avenida '+value.replace('AV.','')* para añadir 'Avenida' al mismo tiempo que eliminamos 'AV.' **(añadir caracteres)**
- *replace(value, /\d/, '')* para eliminar la parte numérica de las celdas **(reemplazar valores)**
- *replace(value, /\D/, '')* para eliminar la parte no numérica de las celdas **(reemplazar valores)**

Podemos usar expresiones regulares (por ejem: *value.replace(regex, "")*

Eliminar espacios blancos del inicio y final de los valores: *Edit cells>Common transforms>Trim leading and trailing whitespace*

Eliminar registros seleccionados (mostrados a la dcha): *All -> Edit rows -> Remove all matching*

rows

Clustering (limpieza basada en similitudes)

A veces, en panel de la izq (filtro), aparece botón '*Cluster*' que abre nueva ventana en la que podemos fusionar (merge) los diferentes grupos escribiendo en todos el mismo nombre como 'nuevo valor de celda'.

Sino, podemos hacerlo con: *edit cells>cluster and edit*.

Repetir proceso para todos los algoritmos cambiando el 'método' y la 'función' (también el 'radius').

Limpiar valores de columna numérica:

- *Facet>numeric facet* (a la izq seleccionamos sólo el tipo '*non-numeric*' para transformar sólo éstos)
- Si a la izq vemos que, erróneamente, los valores están muy dispersos tenemos varias opciones:
  - usar escala logarítmica para corregir: clic en *Change* (submenú izq) y escribir: *value.log( )* (Aceptar si previsualización es correcta).
  - clic y arrastrar extremos del gráfico de la izq para visualizar subconjunto de datos más dispersos sobre los que hacer las transformaciones necesarias
- Hacer las 'transformaciones' necesarias: *edit cells>transform* (por ejem: reemplazar ',' por '.' para decimales)
- *Edit cells>common transforms>to number* (para terminar de convertir los números en valores numéricos

Limpiar fechas

1. convertir valores de columna a texto: *Edit cells -> Common transformations -> To text* (para evitar q haya numeros)
2. convertir valores a fecha: *Edit cells -> Common transformations -> To date*
3. mostrar patrones de fecha: *Facet -> Timeline facet*
4. seleccionar sólo los valores '*non-time*' para transformar sólo éstos
5. extraer sólo el año mediante una expresión regular: *Edit cells -> Transform* y escribir: *value.match(/.*(\d{4}).*/)[0]*
   - el ". *" significa una secuencia de cero o más caracteres (letras, números, símbolos, etc.)
   - el "\ d" indica que estamos buscando a un dígito.
   - el "{4}" muestra que queremos coincidir exactamente 4 cifras.
   - la función *value.match* devuelve una matriz de resultados, de manera que usamos "*[0]*" recuperar sólo la primera coincidencia.
6. convertir estos valores extraidos a fechas: *Edit cells -> Common transformations -> To date*

Si hay fechas con varios formatos: *Edit cells -> Transform* y podemos extraer el valor que nos interese escribiendo:

- *value.toString('yyyy')* obtenemos: 2013
- *value.toString('M')* obtenemos: 1
- *value.toString('MM')* obtenemos: 01
- *value.toString('MMM')* obtenemos: Ene
- *value.toString ('MMMM')* obtenemos: Enero

Podemos usar un sólo código para obtener un sólo formato: *value.toDate('MM/yy','MMM-yy').toString('yyyy-MM')*

Partir/Unir columnas

Separar en distintas columnas: *edit column>split into several columns* (poner separador y el nº de columnas; funciona de izq a dcha). Después podemos renombrar nueva columna: *Edit Column -> Rename this column*

Añadir nuevas columnas

Añadir una nueva columna basada en otra columna:*edit column>add column based on this column*

Poner un nombre a la nueva columna y, por ejem, escribir:

- *cells("direccion").value+" "+cells("ciudad").value*
- *value.startsWith("-")* ó *not(value.startsWith("-"))*
- *value[1,5]* y la nueva columna contendrá los valores 2º al 5º (el 1º es 0)
- *value.substring(6)* para extraer a partir del caracter nº 7
- *value.substring(3,5)* para extraer los caracteres entre el 4º y 6º
- *value.facetCount("value", "keyword")* para crear nueva columna con el nº de repeticiones de 'keyword'
- *cells["A"].value / cells["C"].value* para nueva columna con resultados de división de 2 valores

Geocodificación

Para geocodificar nombres y direcciones (convertir una dirección a sus coordenadas basándose en Google Maps)

Google limita a 2000 solicitudes/día (nejor sólo seleccionar algunas filas)

1. seleccionar unas pocas filas
2. en la columna que contiene nombres de empresas o direcciones: *Edit column>Add column by fetching URLs* , escribir nombre para nueva columna e introducir: "*http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")*
3. En columna creada: *Edit column>Add column based on this column* e introducir:*with(value.parseJson().results[0].geometry.location, pair, pair.lat +", " + pair.lng)*

Exportar datos a excel

En menú superior dcha: Export>Excel