Preparación de los datos – Limpiar los datos con OpenRefine

En este video, les mostraré cómo usar OpenRefine y sus excelentes funciones que tiene para limpiar conjuntos de datos.

Vayan a openrefine.org y hagan clic en “Download” (descargar). Pueden descargar la versión que se aplica a su sistema operativo. Hay una versión para Windows, una versión para Mac y una versión para Linux. Cuando descarguen OpenRefine, pueden abrirlo, tiene este logotipo en forma de diamante. Si están en Windows, se abrirá una pantalla en negro. No se preocupen. Dejen la pantalla en negro abierta, es importante para ejecutar el programa.

Usaré el gran tutorial de Sarah Cohen sobre cómo limpiar un conjunto de datos usando OpenRefine. Utilizamos los mismos conjuntos de datos en la herramienta de preparación de datos en la nube y utilizaremos un enfoque similar pero diferente para limpiar este conjunto de datos con OpenRefine. Y recuerden, este es el conjunto de datos que Sarah, que entonces estaba en el New York Times, usó en su presentación en la Conferencia de Reportería Asistida por Computador (Computer Assisted Reporting Conference), la conferencia NICAR en Denver, en 2016. Asistí a una clase y fue una gran clase, un gran taller. Y ella construyó este conjunto de datos a partir de los planes de salud de Medicaid en el estado de Nueva York mientras intentaba decidía sobre cuál compañía haría una investigación. Y construyó, cada página en la hoja de cálculo fue un mes, y construyó esta hoja de cálculo grande, grande, grande con todos los archivos juntos en uno. Y luego usó OpenRefine para limpiarlo, especialmente las expresiones regulares. Usaremos algunas hoy, pero no las usaremos todo el tiempo.

Así que esta es la hoja de cálculo y lo que realmente queremos es obtener una columna con los nombres de los planes, los condados, la inscripción total, una columna de mes y una columna de año. Algo que se parece a esto. Plan, condado, inscripción, mes y año. Y para hacer eso, usaremos OpenRefine. También pueden ver el tutorial de Sarah aquí, es genial. Tiene algunas de los recursos más avanzados de OpenRefine. Vamos a familiarizarnos con esto.

Esta es la interfaz, por lo que cuando abren OpenRefine, OpenRefine se ejecuta en su computador. No funciona en la web. Todo lo que hacen se queda en sus computadores. No tienen que estar conectados a internet para usar OpenRefine. Aunque usa su navegador, tengan en cuenta que usa una dirección diferente. Esta dirección apunta a su máquina local. Todo se ejecuta localmente y noten que cuando abren OpenRefine, tiene una interfaz que les permite crear un proyecto, por lo que pueden elegir entre una variedad de fuentes aquí, incluso datos de Google. En el portapapeles pueden pegar datos, pueden poner URL, pueden obtener archivos de un computador, pueden abrir proyectos existentes, pueden importar proyectos de otros archivos. También hay configuraciones de idioma.

Así que sigamos adelante y creemos un proyecto. Este proyecto lo creamos a partir del archivo Excel que Sarah creó. Entonces elegiremos el archivo aquí. Estos datos provienen de este

computador y OpenRefine abre casi todos los formatos de archivos de bases de datos que existen. Así que no deben preocuparte por eso. Archivos DSV, CSV, SV, Excel, JSON, XML, RDF e incluso documentos de datos de Google son admitidos. Por lo tanto, es bastante amplio, vale la pena intentarlo.

Así que abramos este. Luego hago clic en “next” (siguiente) y me da una vista previa de todo aquí. Le daré un nombre a este proyecto, “Planes de Medicaid en Nueva York”. Y les ofrece una vista previa para que puedan ver si está bien, si el código es correcto. Aquí les solicita “parse data as” (analizar datos como) un archivo de Excel por lo que ya ha identificado que es un archivo de Excel. Así que vamos a dejarlo así. Hay otros formatos aquí. Lo que desmarcaremos aquí es “analizar la primera línea como encabezados de columna”. A veces ese es el caso, pero este no es el caso aquí. Como podemos ver, la primera fila no es la fila de encabezado. Por lo tanto, no analizaremos la primera línea como una fila de encabezado, y luego crearemos columnas aquí. Entonces, todo se ve bien.

Vamos a construir el proyecto ahora y familiarizarnos con la interfaz de OpenRefine. Todo lo que hacen en OpenRefine sucede en esta interfaz. Tienen aquí en la izquierda un filtro y un menú de facetas, que les explicaré pronto. Y aquí a la derecha, siempre tendrán en negrita la cantidad de registros en los que está trabajando en el momento, ¿verdad? También pueden cambiar a filas y registros. Estaremos usando filas. Estas son filas que normalmente ven en archivos como este. Entonces, por ejemplo, si vemos aquí, es 6.664, que es el mismo número de filas que estamos viendo aquí. También pueden ver que el número de filas se mostrará cuando lo configure en 50. Pueden abrir proyectos desde aquí, pueden en cualquier momento exportar su proyecto, lo que sea que estén viendo aquí. Esto es otra cosa, puede ser un archivo de Excel, una tabla HTML o un archivo CSV o cualquier cosa. Y también pueden navegar por el archivo aquí, siguiente, primera y última página. Y todas las opciones en OpenRefine están aquí en la parte superior de las columnas, por lo que todo lo que necesiten hacer lo harán para una columna específica, o pueden aplicarlo a todas las columnas, si eso es lo que necesitan.

Entonces, lo primero que haremos aquí, ya que queremos convertir este desordenado conjunto de datos en un conjunto de datos organizado como este, lo primero que haremos es crear una columna de año aquí. Y para crear la columna del año, necesitaremos extraer este número aquí de esta columna y luego agregarlo a una nueva columna. Así que la manera en la que lo vamos a hacer es haciendo clic aquí en esta flecha. Hacemos clic en “edit column” (editar columna) y luego “add column based on this column” (agregar columna basada en esta columna). Y luego hay una nueva ventana donde pueden hacer transformaciones aquí. Entonces vamos a usar una expresión regular. Voy a nombrar esta nueva columna “año” y aquí simplemente escribiré “find” (encontrar) para que encuentre esto que quiero aquí.

Lo que estoy usando aquí es básicamente comenzar la expresión regular con una barra inclinada / y terminar la expresión regular con una barra inclinada /. Lo que estoy escribiendo aquí es algo así como encuentre para mí el número de cuatro dígitos que está en el valor, y

luego devuélvamelo, que es esto. Si eliminamos esto, nos da una lista. En la programación de computadores, por lo general, el primer elemento de una lista es la posición cero, comienza con cero. Entonces estoy diciendo “deme la posición cero de la lista de resultados que está mostrando”. Así que esta es solo una expresión regular para extraer el año de esta columna y luego hago clic en “OK” y luego crea una nueva columna con el año. Ahora quiero extraer el mes y ponerlo aquí, pero el mes es un poco más complicado. El mes será “edit column”, “add column based on this column”, y lo que voy a usar aquí es un poco más complicado, pero es bastante fácil de entender. Así que también comienzo con la barra inclinada / y cierro con la barra inclinada /. Miren aquí, es NYS, es una palabra y luego dígitos, entonces NYS, una palabra, y luego dígitos. Lo que agregué aquí en el medio, que son estos corchetes, y luego tengo una coma y un espacio, corchetes, una coma y un espacio, lo que le estoy diciendo a OpenRefine es ‘mire, busque NYS y no me importa si después de NYS viene una coma o un espacio tantas veces como sea’. Esto es lo que el más (+) significa. Por lo tanto, no me importa si hay una coma aquí, o un espacio, o un espacio doble debido a los errores tipográficos en este conjunto de datos: pueden encontrar todos estos casos. Realmente no importa si es una coma o un espacio varias veces. Traiga la palabra que aparece aquí, justo en el medio. Es por eso que estoy poniendo esto entre paréntesis, esto es para capturar un grupo. Y puse esto aquí para que podamos entender mejor esta expresión, ¿correcto?

Entonces lo borramos para entender lo que está pasando aquí. Lo que está sucediendo aquí es que estoy capturando aquí en verde solo la palabra “january” (enero), ¿verdad? Los paréntesis en la expresión regular significa “capturar lo que hay dentro”. Así que estoy capturando la palabra que está entre este espacio o una coma que viene antes de cuatro dígitos, que aparece después de un espacio o una coma después de NYS. Entonces estoy tomando la palabra en el medio. Eso es lo que estoy haciendo aquí en OpenRefine y luego este captura aquí, enero, que es exactamente lo que necesito y cambiaré el nombre de esta columna a “mes”.

Ahora necesito completar todos los valores aquí, porque encontró enero en 2009, pero también tengo 2010 aquí. Si van al último, verán que hay otros años aquí, noviembre de 2009. Y así sucesivamente para otros meses y años también. Como pueden ver, aquí está diciembre de 2013. Por lo tanto, debemos completar esto para que obtenga todos los valores de mes y año. Así que hagamos esto y es muy, muy fácil hacerlo en OpenRefine. Simplemente hagan clic aquí en el mes, por ejemplo. “Edit cells” (Editar celdas) y luego seleccionen “Fill down” (Rellenar), y listo. Entonces hacen lo mismo aquí. “Edit cells” y luego seleccionen “Fill down” y lo mismo pasa con el año.

Ahora comencemos a limpiar y para eso filtraremos, lo cual es realmente genial porque luego pueden eliminar según el filtrado. Entonces, lo primero que vamos a hacer es eliminar las filas que tienen “total” aquí. Vamos a crear un filtro de texto aquí y escribimos “total”. Él me mostrará aquí todas las filas, 74, que tienen la palabra “total”. Y luego hacemos clic en “all” (todos), luego “edit rows” (editar filas), y ahora “remove all matching rows” (eliminar todas las filas coincidentes) y luego eliminar 74 filas. Luego limpiamos el filtro y nos muestra nuevamente.

Entonces, vamos a hacer esto un par de veces más para tener una mejor limpieza, así que vamos a eliminar las celdas en blanco, por ejemplo, para que hagamos “facets”, “customized facets”, y luego “facet by blank”. Por lo tanto, las facetas son formas de identificar categorías dentro de su columna. Entonces, al crear una faceta, están tratando de identificar celdas que tienen el mismo valor, así que vamos a agruparlas para que puedan ver las categorías. Vamos a separar las que están en blanco y luego hagámoslo por texto para que puedan ver mejor a qué me refiero. Así que voy a crear una “facet by blank”, lo que significa una cadena nula o vacía y luego me dice que 842 filas aquí están en blanco, lo que significa que están vacías o no hay un valor final. Entonces hagamos clic aquí y me da las 842 filas coincidentes y quiero eliminar todas estas filas porque no son necesarias para el conjunto de datos final. Simplemente eliminaré todas estas filas y borraré la faceta aquí.

Una vez más, tengo el condado y las inscripciones aquí. Voy a continuar y crearé filtros para eliminar esto y las inscripciones. Eliminaré todas las filas coincidentes aquí. Y ahora, lo que quiero hacer es completar los nombres de los planes, ¿verdad? Porque todos pertenecen a los mismos planes aquí. Entonces, lo que voy a hacer es “edit cells” (editar celdas), también “fill down” aquí y eliminar el total aquí de esta columna también. Todo bien. Eliminar todas las filas coincidentes. Ya se ve mucho, mejor, pero todavía hay algunos que podemos hacer faceta en blanco aquí también. Vean eso. Pueden eliminar todo eso. Y luego limpiarlo. Muy bien, se ve mucho, mucho, mucho mejor. Podemos renombrar esta columna ahora, “edit column”. Nombren esta columna, quiero llamarla “Planes”. A esta la llamaremos “condado”. Y esta será “total”.

Ahora, otra cosa que podrían hacer es usar una herramienta muy, muy poderosa. Podemos crear una faceta de texto aquí para ver cuáles son los planes y luego podemos clasificarlos. También podemos ver los condados aquí, creando una faceta de texto para ver qué condados se enumeran aquí y con qué frecuencia se enumeran los condados. Entonces pueden ver los condados de Nueva York que se enumeran con mayor frecuencia.

También pueden hacer esto para los planes. Pueden verificar si los planes están escritos correctamente o pueden crearlos, vean aquí, “Eddy SeniorCare”, “seniorcare” está escrito junto y debería estar separado. Hay una función muy poderosa en OpenRefine llamada cluster. Y cuando hacen clic en el cluster, este trata de detectar errores ortográficos y utiliza diferentes metodologías para hacerlo, y les dejaré explorar. Pueden explorar las metodologías aquí. También hay métodos y funciones clave, pero lo que hace es tratar de encontrar similitudes en los nombres que se escribieron casi de la misma manera. Entonces, Eddy SeniorCare, por ejemplo, este, y pueden navegar por este cluster para ver qué está pasando aquí. Y vean qué filas y su significado, ¿verdad? Todos estos están en Eddy Seniorcare. Están todos juntos, pero estamos viendo aquellos que son muy similares. Y queremos que corrijan la ortografía. Entonces, por ejemplo, en este tenemos a Eddy Senior Care y Eddy Seniorcare, escrito junto. El que queremos es el Eddy Senior Care escrito separado, así que fusionemos este y luego “merge selected and re-cluster” (fusionar lo seleccionado y volver a agrupar) y usémoslo, entonces 171 celdas que fueron editadas aquí, así que en una fracción de segundo editó 171

celdas corrigiendo el error tipográfico. Entonces pueden cambiar las metodologías aquí y ver cuál funcionaría mejor para ustedes, pero esta también es una herramienta poderosa para corregir errores tipográficos.

Entonces, de ahora en adelante, si esto les luce bien, pueden exportar la tabla a cualquier formato que deseen, CSV o Excel o TSV, incluso pueden exportar el proyecto para abrirlo en otras versiones de OpenRefine. Entonces exportaré esto como un CSV. Y luego me pide que descargue el archivo CSV, como cualquier archivo que descarguen de internet. Así que aquí está, el archivo CSV final. Y vean que está listo para usar.

Esta es una introducción rápida a OpenRefine. Continúen y abran openrefine.org/download y descarguen la versión que sea mejor para su sistema operativo y comiencen a limpiar. Y cualquier pregunta que tengan, simplemente publíquenla en los foros y nos veremos allí.