Taller 2: Acceder a datos desde diferentes fuentes de información

Una de las primeras tareas en el proceso de minería de texto es seleccionar las fuentes de datos que permite encontrar los patrones dentro del texto. Dichas fuentes pueden tener diferentes formatos. Esta práctica tiene como meta acceder a fuentes de diferentes formatos

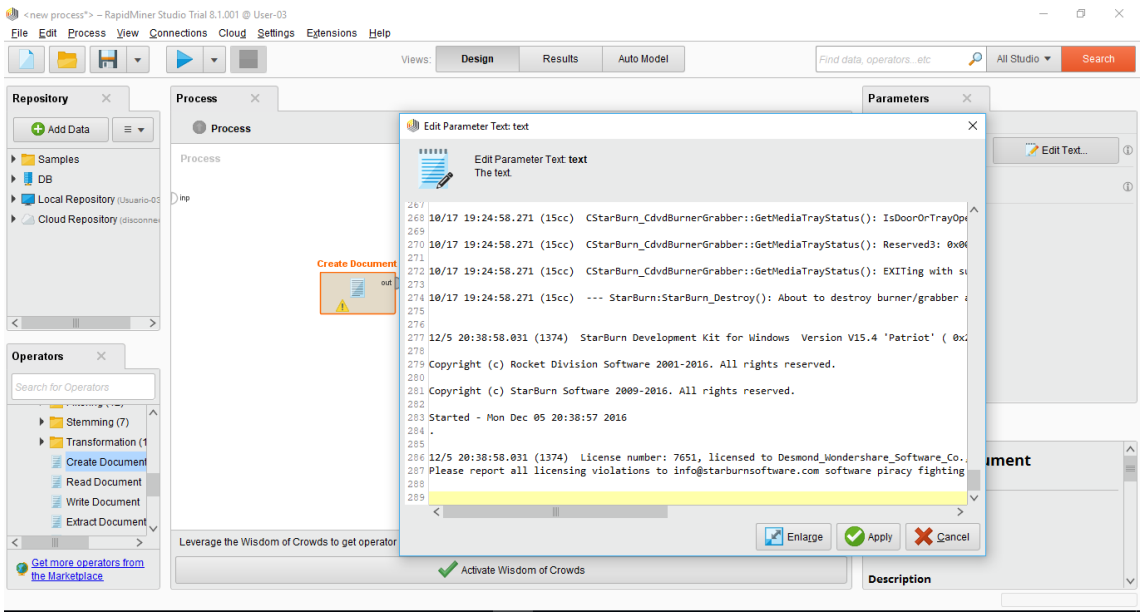
Lo primero es verificar que los componentes Text Processing y Web Mining están instalados dentro de la herramienta. Para comprobar aquello es necesario acceder al menú Extensions, Market Place (Update and Extensions). Si aún no están instalados seleccionar los componentes para ejecutar la instalación.

La extensión Text Processing proporciona todos los operadores de minería de textos necesarios para extrapolar completamente y revisar estadísticamente los datos de tipo texto dentro de un corpus (por ejemplo, reduciendo las palabras a su raíz, agregando el filtro de stopword).

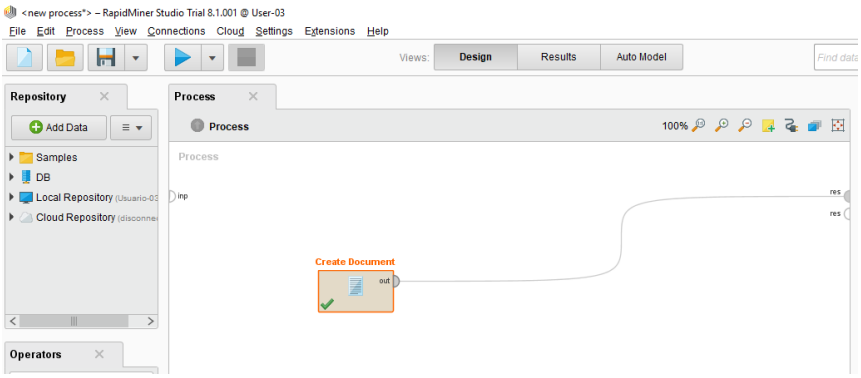
Existen diferentes formas de acceder a repositorios que proveen información textual:

Manera Indirecta:

Usando el operador *Create Document* es posible crear un documento nuevo como fuente de datos para minería de texto. Para comprobar su uso, agregue dicho componente a la interface de procesos y ejecute el resultado. El archivo que puede usar para copiar el contenido dentro del componente se llama ejemplo.txt. Para ello use la opción Edit Text… en la pestaña de parámetros



Para observar el documento conecte el método out al resultado del proceso res y ejecute.

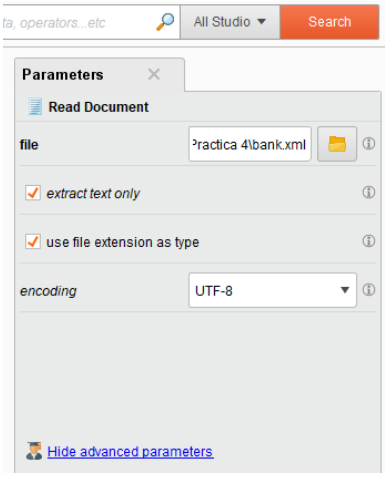


Manera Directa:

Para leer diferentes fuentes de texto de manera directa es posible usar el componente Read Document.

Opcion 1:

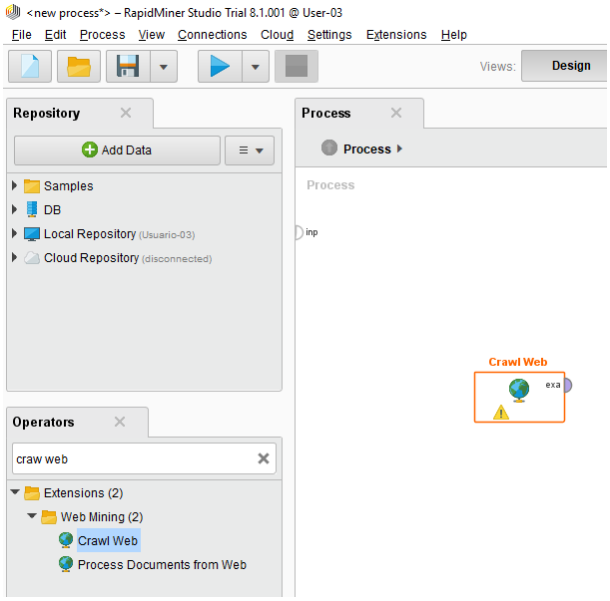
Lo primero es leer un archivo XML y extraer del archivo la información textual. Para esta práctica se propone usar el archivo bank.xml como ejemplo. Note que en la sección parámetros es posible configurar que únicamente se extraiga el texto del documento usando la opción extract text only. Esta opción no toma en cuenta las etiquetas XML o en el caso de una página web las etiquetas HTML



Opción 2

Se puede rastrear una página web, para ello se debe tener instalada la extensión Web Mining. La extensión Web Mining funciona en complemento con la extensión de procesamiento de texto; por lo tanto, esto necesita ser instalado en conjunto. La extensión Web Mining proporciona operadores (como Get Pages) en los que puede recuperar información en línea, por ejemplo, mediante técnicas de *Web Scraping*.

Una vez que haya descargado la extensión Web Mining, abra esta carpeta en la sección de Operadores y luego seleccione y arrastre la opción Crawl Web a la sección de Proceso.

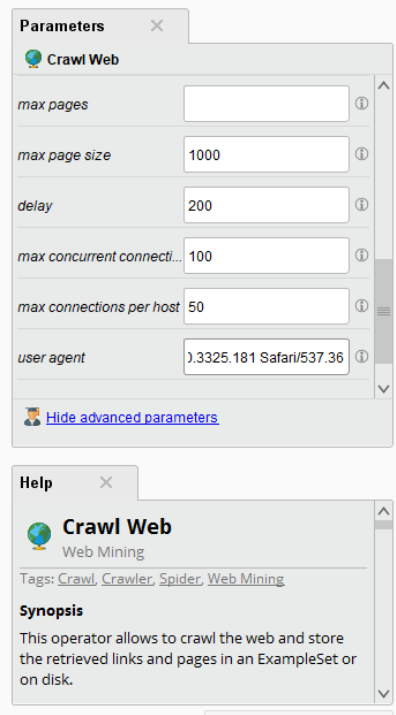


Una vez que haya hecho esto, debe elegir un sitio web para rastrear. Por lo tanto, copiamos y pegamos la url en el cuadro respectivo en el lado derecho de la pantalla, debajo de la pestaña de parámetros. Una opción es la url: http://www.ucuenca.edu.ec

Luego, debe seleccionar un directorio de salida para que RapidMiner guarde los archivos. Puede seleccionar cualquier carpeta por ejemplo "tmp". Luego, si desea seleccionar una extensión de archivo, puede elegir .txt. La herramienta guardará los archivos que rastrea como archivos de texto. La profundidad máxima es la cantidad de enlaces consecutivos que seguirá el rastreo, puede elegir el valor predeterminado 2. El campo domain permite saber si el rastreo permanecerá en el mismo servidor o si permitirá rastrear todo el sitio web, por omisión el valor es la web.

Se puede establecer además los hilos máximos, que es el número de núcleos de CPU que utilizará el rastreo, un ejemplo es colocarlo en un valor de 4 para acelerar el rastreo. Luego, cambié el agente de usuario por el del navegador que está usando. Para hacerlo, vaya al siguiente enlace para descubrir el agente usado por su navegador

http://whatsmyuseragent.com/ y copie y pegue su agente de usuario en el cuadro.



Es posible configurar algunas reglas de rastreo. Para ello, haga clic en el botón al lado de las reglas de rastreo que dice "Edit List"

Como se puede ver en el dialogo se pueden agregar reglas de rastreo. Algunos ejemplos de reglas pueden ser:

- following_link_wtith_matching_url: Esta regla permite seguir cualquier enlace con la palabra test en la url. En este caso en el campo rule_value puede colocar .+test+. El símbolo + indica que cualquier cantidad de caracteres antes y después de test puede estar en la url.
- store_with_matching_url: Esta regla guarda solo las páginas que tienen la palabra test en la url. El mismo valor que el ejemplo anterior puede colocarse en el campo rule_value

Opción 3

Otra opción es leer un conjunto de datos desde Excel o una Base de datos especialmente cuando se conoce a priori los textos que deseamos leer de estas fuentes. Para ello usaremos el componente Read Excel y lo agregaremos a la interface de procesos. Para la práctica usar como ejemplo el archivo ejemplo.xlsx. Únicamente seleccionar las casillas desde C8:V33.

Puesto que la mayoría de técnicas de procesamiento de texto requieren documentos es posible transformar los datos a documentos.

Para ello se puede usar el conector Data To Documents. Así se la hoja tiene 10 filas entonces se creará 10 documentos con información de las columnas contenidas en dicho tabla.

Opción 4

La última opción de esta práctica es leer texto desde múltiples archivos almacenados en el computador en un grupo de directorios. Esta opción puede ser útil para ejecutar web crawling, por ejemplo colocando las diferentes categorías de noticias en un directorio diferente dependiendo de los textos encontrados. Para ejecutar este proceso se puede usar el conector Process Documents from Files

En el campo text directories puede configurar los directorios desde donde es posible extraer la información de los archivos que coincidan con la extensión propuesta

Ejercicio Propuesto

Usando la opción 2 para rastrear una página web, usar las reglas de rastreo de forma que sea posible rastrear otras páginas adicionales que cumplan algunos criterios adicionales que cumplan las reglas especificadas