Taller 1: Acceso a una fuente de datos

En esta práctica se ejecutará un proceso sencillo de acceso a una fuente de datos usando una herramienta para usuario final

Desde el siguiente enlace, bajar la versión de RapidMiner correspondiente a su sistema operativo

https://my.rapidminer.com/nexus/account/index.html#downloads

El objetivo de la práctica es leer un archivo que contiene un listado de palabras vacías (stop words). Este tipo de palabras son aquellas que al procesarlas no aportan mayor significado. Normalmente estas palabras son usadas para crear un modelo de minería de texto y es uno de los primeros pasos en la tarea de pre-procesamiento

Lo primero que es necesario determinar es si se tiene instalado el componente **Text Processing**. Para ello es necesario verificar en el menú Extension, MarketPlace (Update and Extensions). La extensión Text Processing admite varios formatos de texto, incluidos texto sin formato, HTML o PDF, así como otras fuentes de datos.

Si el paquete no está instalado se puede elegir instalar el componente luego de ejecutar el proceso de búsqueda

Una vez instalado el componente escogerlo y usar el operador Read Document el cual será usado para leer el documento. En la sección de Parameters se escoge la ruta del archivo que pretendemos usar para generar la lista

Desde el componente se une la salida (out) con el resultado del proceso (res) y usando el botón de ejecutar se comprueba que el archivo sea leído correctamente. El nombre del archivo a usar es stopword.txt

Otro de los componentes muy usado al momento de hacer minería es la extensión **Web Mining**. Esta extensión proporciona acceso a varias fuentes de Internet como páginas web, fuentes RSS y servicios web. Además de los operadores para acceder a esas fuentes de datos, la extensión también proporciona operadores específicos para el manejo y la transformación del contenido de las páginas web para prepararlas para su posterior procesamiento.

Una vez instalado el paquete use el operador Get Page para conseguir una página web via HTTP. Proporcione los parámetros necesarios para la lectura de la página. Para este ejemplo use la URL de la página web de la Universidad de Cuenca. El mismo procedimiento usado para visualizar el contenido del archivo en el ejemplo anterior debe ser usado en este caso.

Para comprobar el funcionamiento de este componente puede deshabilitar momentáneamente el operador Read Document dando un clic derecho sobre el componente y escogiendo la opción Enable Operator.

Por último pruebe el operador Read RSS Feed el cual permite leer un archivo RSS desde una URL. Dicho operador es parte además del componente Web Mining. Use el RSS de deportes perteneciente a diario el comercio: http://www.elcomercio.com/rss/deportes