Práctica 5

En la práctica previa se crearon los pasos iniciales de la construcción de un repositorio basado en texto; En este ejercicio se ejecuta la segunda actividad que consiste en recopilar los detalles relevantes principales de cada discurso, extrayendo información clave para nuevos atributos como

- nombre del presidente,
- fecha del discurso y
- contenido del discurso.

Esta información extraída se debe almacenar dentro de un repositorio de base de datos una vez que se recupere con éxito.

Paso A: Ahora que la hoja de cálculo se creó en la práctica previa es posible importar y explotar activamente la estructura HTML de los discursos inaugurales.

El primer proceso es acceder a la hoja de cálculo utilizando el operador Read Excel. Haga clic en el botón Import Configuration Wizard seleccione el archivo.xls y haga clic en el botón [Siguiente]. Hay cuatro pasos dentro de este proceso; presione [Siguiente] hasta llegar al cuarto paso y haga clic en [Finalizar].

Paso B: ahora que tiene una conexión con la hoja de cálculo que contiene las URL, use el operador Get Pages para recuperar los discursos utilizando la lista de direcciones HTML. Este operador no debe confundirse con el operador Get Page, que recupera solo una página. Obtener páginas funciona examinando el atributo Link = URL y utilizando el valor de los datos (es decir, los enlaces URL de la Web para los discursos) dentro de este atributo para determinar el conjunto de ejemplos y, por lo tanto, la extracción de la información adecuada.

Por lo tanto, los propios datos de la hoja de cálculo contenidos en Link = URL determinan y, por lo tanto, controlan la información que se recupera para futuros análisis cuando se usa el operador Get Pages. Funciona dentro de un bucle recorriendo cada fila del atributo Link = URL y seleccionando el siguiente valor de ese atributo hasta que no haya más disponible para ser llamado. Asegúrese de que el atributo del enlace sea Link = URL

Paso C: ahora necesita extraer los elementos clave del nombre del presidente, la fecha del discurso y el discurso mismo del documento original por cada discurso inaugural. Seleccione el operador Processs documents from Data y tenga los siguientes procesos dentro de su ventana Creación de vectores

Extract Information operator: Use una expresión regular para extraer

- año del discurso inaugural. Etiquete este valor como Fecha Inauguración.
- nombre del presidente. Etiquete este valor como Presidente
- el discurso inaugural. Etiquete este valor como Texto. Una opción podría ser usar el operador CUT para extraer esta información

Paso D: El último paso en este proceso es almacenar la información usando el operador Repository Access:Store, en este caso en un repositorio de RapidMiner llamado Discursos

El siguiente requisito dentro de este segundo elemento de procedimiento es crear un proceso que escriba en la base de datos MySQL denominada, discursosinaugurales. Esto permite acceder, manipular y consultar los datos usando SQL en las siguientes prácticas.

Paso A: use el operador Repository Access: Retrieve para acceder al repositorio almacenado Discursos.

Paso B: el siguiente operador que necesita es el operador Rename, ya que desea cambiar los nombres de dos atributos. Estos cambios de nombre mejoran la legibilidad de su salida.

Paso C: de todos los atributos en el repositorio, desea almacenar solo tres: fecha inaugural, presidente y discurso. Para hacer esto, usa el operador Select Atributtes y seleccione su subconjunto.

Paso D: Finalmente, para almacenar esta información en su base de datos MySQL, debe conectarse utilizando el operador Write Database. Dentro de esto, establezca una configuración de conexión a la base de datos discursosinaugurales.

Finalmente ingrese un nombre de tabla tblDiscursos y seleccione la opción overwrite first, append then.

Una vez que ejecute su proceso, esa tabla se creará y completará con todos los discursos indexados disponibles.

NOTA: Crea una base de datos MySQL llamada discusosinaugurales. No cree tablas, ya que se generarán automáticamente en los procesos de RapidMiner