Práctica 4

Los objetivos de esta práctica es presentar los fundamentos de la construcción de un repositorio basado en texto;

Para los objetivos de la práctica se debe tener instalados los siguientes complementos dentro de su copia de RapidMiner: Information Extraction, Text Processing y Web Mining

Proceso de construcción del corpus

Descripción general

El objetivo es construir un corpus que pueda usarse para etapas posteriores de la minería de texto. Para ello, debe obtener una lista de enlaces de direcciones URL en funcionamiento desde donde sea posible extraer contenido, luego analizarlo y finalmente producir informes visuales.

Esta primera parte de la práctica se concentra en el paso inicial: crear una lista de URLs que se puedan descargar. Para llevar a cabo esta tarea, necesita que se recopile la etiqueta de referencia única de cada discurso (etiquetada como PID en la estructura de la URL)

La fuente que se usará es el sitio web "The American Presidency Project", en particular la página web que contiene índices de todos los discursos inaugurales de los presidentes estadounidenses: http://www.presidency.ucsb.edu/inaugurals.php

La primera tarea antes de ejecutar el proceso de minería de texto es comprender el corpus. Esta página web provee información sobre el número de discursos, lo cual da una apreciación del contexto del corpus. Esta página web también enumera el número aproximado de palabras de cada discurso; esta información es valiosa, ya que le permite medir la cantidad de tokens para su análisis. Para comprender la estructura del contenido HTML, se debe seleccionar la opción *Ver código fuente de la página*. Este conocimiento es esencial para el modelo de extracción de texto, ya que informa de los tipos de métodos de recuperación que serán necesarios para recopilar los fragmentos de datos para el análisis. Este proceso se puede inferir al ver la estructura HTML que es donde se encuentran los elementos de datos a extraer.

NOTA: El lenguaje en todos los discursos presidenciales es el inglés, aunque la terminología y la construcción verbal cambian en los diferentes períodos. Todos los discursos se guardan, desde el 30 de abril de 1789 (George Washington) hasta el día de hoy (el actual presidente estadounidense).

Creación del Repositorio

Descargar información de Internet

El primer procedimiento para construir el repositorio de esta práctica está relacionado con la recopilación de una lista de URLs con identificadores únicos (representados por la variable PID en la dirección URL) que define claramente cada discurso.

Para lograr esto, se requieren varios operadores de la herramienta RapidMiner.

Get Page

Este operador permite descubrir cuál es el rango de identificadores únicos de los discursos, para ello debe descargar y extraer información de http://www.presidency.ucsb.edu/inaugurals.php,

que contiene estos valores dentro del código de contenido en HTML. El operador Get Page se puede usar para recuperar estos valores, cuyo contenido HTML contiene los identificadores necesarios que hacen referencia a todos los discursos inaugurales, por ejemplo:

```
<a
href="http://www.presidency.ucsb.edu/ws/index.php?pid=25801
">
```

En resumen, el operador Get Page recupera la página web a través de HTTP y genera la información como un documento de código fuente HTML. Este documento puede ser cortado y extraído para obtener los números de referencia que desea recuperar.

Una vez recuperada la página se recomienda usar los siguientes operadores:

Process Documents

El elemento donde se puede extraer el número de referencia es PID = 25801 (en el caso del ejemplo mostrado arriba). Para lograr esta extracción, se recomienda estudiar los operadores descritos a continuación. La idea es crear un subproceso dentro de este operador.

Cut Document

El operador Cut Document divide el código fuente HTML en segmentos utilizando criterios específicos que pueden basarse en una gama de mecanismos de tipo de consulta que incluyen:

- String Matching: Especifica un comienzo y un final para una búsqueda de cadenas y extrae todo lo que se encuentre entre esos dos parámetros.
- Regular Region: Este tipo de consulta es esencialmente una forma fácil de agregar 2 expresiones regulares para estipular una etiqueta de inicio, y una etiqueta de finalización para recuperar lo que está entre esos dos parámetros.
- Regular Expression: son expresiones usadas para unir un grupo y extraer la información apropiada. Requieren configuración detallada y conocimiento de varias palabras de comando y funciones de símbolos. Tiende a ser más flexibles que otros tipos de consultas

Para una referencia completa de estos términos, seleccione el menú desplegable query type dentro de este operador.

Extract Information

Este operador puede ser usado para extraer la información requerida, por ejemplo cada número de referencia de todos los discursos inaugurales.

Note que la diferencia principal entre el operador Cut Document y el operador Extract information es que el primer mecanismo corta un documento, como un discurso, usando una etiqueta de inicio y final, generalmente estipulada por expresiones regulares, mientras que el segundo operador extrae información dentro de un texto mejor estructurado. Por lo tanto, a menudo se puede usar estos dos operadores en conjunto. Uno como subproceso del otro, por ejemplo para cortar un documento primero usando el operador Cut Documento y luego como un subproceso usar el operador Extract information para extraer información de esa porción de texto.

Para esta práctica identifique el mecanismo más adecuado para extraer el identificador PID de cada discurso.

El siguiente paso es crear un atributo que contenga la dirección de enlace para cada discurso que desea rastrear y analizar. Seleccione el operador Generate Attributes en RapidMiner. El contenido del atributo debe ser http://www.presidency.ucsb.edu/ws/index.php?pid. El atributo debería llamarse como ENLACE.

La siguiente tarea es unir la información extraída previamente, esto es, el número de referencia PID que se encuentra dentro de la URL con el enlace HTTP denominado ENLACE para crear una dirección URL completa para cada discurso y almacenar esta lista dentro de una hoja de cálculo.

Estudie el operador Generate Concatenation para unir estos dos atributos ENLACE y URL, con un signo =, como se puede consultar en el siguiente ejemplo de salida: http://www.presidency.ucsb.edu/ws/index.php?pid=25824.

Finalmente, use el operador Write Excel para almacenar las direcciones URL de los discursos generadas dentro de una hoja de cálculo; debe tenerse en cuenta que debe asegurarse de que el formato de archivo sea XLS.