Práctica 6

En la práctica previa se creó una tabla completamente poblada almacenada dentro de una base de datos MySQL que contiene tres atributos: Discurso, Fecha Inaugural y Presidente. La meta de los talleres previos es conocer las diferencias entre las expresiones regulares, la coincidencia de cadenas y las regiones regulares. También debería estar familiarizado ahora con las diferencias entre los operadores Information Extraction y Cut Document.

El objetivo de esta práctica es Analizar el Corpus ahora que se ha reunido toda la información, para ejecutar algún proceso. Hay una gran cantidad de formas en la que es posible buscar y, posteriormente, analizar el contenido. En el caso de este ejemplo, se escribirá una consulta estándar para recuperar el discurso del presidente Barack Obama en 2009 y luego utilizar los procesos de minería de textos para realizar diversas funciones, incluyendo la aplicación de tokens, diccionarios de palabras vacías y n-grams. Estos resultados se almacenarán en una hoja de cálculo y luego se resumirán en las palabras más frecuentemente utilizadas, que luego se mostrarán utilizando un método de visualización llamado nubes de palabras. Este método de visualización será el objetivo en una práctica posterior

Sección uno: Ejecutado minería sobre el repositorio: Frecuencia de palabras

Paso A: el primer paso es consultar y recuperar información relevante de la tabla tbldiscursos usando el operador Read Database

Haga clic en el botón [Build SQL Query] una vez que haya establecido una conexión denominada DiscursosInauguralesPresidente.

Para extraer el discurso del presidente Barack Obama en 2009, ejecute una consulta en SQL

```
SELECT ‘Discurso‘
FROM ‘tbldiscursos‘
WHERE tbldiscursos.FechaInaugural like ’%2009%’ AND
Presidente like ’%Obama%’
```

Paso B: una vez ejecutada la consulta, el discurso debe extraerse y transformarse en tokens; esto se logra utilizando el operador Process Document from data. En el parámetro vector creation se debe seleccionar Term Ocurrencies para garantizar que se calcule el número de tokens (palabras) recurrentes. Se requieren varios procesos ejecutados como subprocesos, para ejecutar el preprocesamiento

Como los datos son de origen HTML, se debe eliminar las etiquetas HTML, como <p>, etc., de modo que puede incluir el operador Extract Content en la ventana de documentos de proceso. A partir de esto, se romperá el discurso en palabras individuales (también llamadas tokens) usando el operador Tokenize.

Hay palabras vacías estándar en inglés como “the” y “is” que no quiere que se complique su análisis; para eliminar automáticamente estas palabras, incluya un operador de diccionario llamado Filter Stopwords (English)

Finalmente, se recomienda utilizar un último operador, que permita crear una lista personalizada de palabras que puede ser usada para eliminar los tokens que saturan los resultados y no tienen ningún beneficio en su inclusión. Este diccionario personalizado se crea a través de un editor de texto cualquiera (ej. Bloc de notas de Windows) y se guarda como una extensión .txt. Averigüe el operado necesario para ejecutar esta acción. En este caso, el diccionario personalizado de stopword se guarda como discursostop.txt

Un ejemplo de esta lista de palabras puede ser encontrado a continuación (estas palabras se excluyen de la salida del análisis). Simplemente enumere las palabras que desea excluir, una línea tras otra. Configure apropiadamente el operador seleccionado para este fin.

Paso C: Para generar los resultados del preprocesamiento, use el operador de Write Excel para escribir en una hoja de cálculo llamada analisissalida.xlsx. Los datos se representan con las palabras en la Fila 1 y la frecuencia de sus ocurrencias en la Fila 2. Para el análisis con nubes de palabras, éstas deben copiarse y transponerse en una nueva hoja de trabajo de modo que las palabras estén en la Columna A (Token etiquetado) y los valores de frecuencia en la Columna B (Cantidad etiquetada), como se hace referencia en la Figura 1.41.

Este informe generado se transformará en una nube de palabras en una próxima práctica

Sección dos: Ejecutado minería sobre el repositorio: Frecuencia de N-grams

La enumeración de palabras por su frecuencia de repetición, como se explicó en la sección previa, proporciona un resumen básico de temas mediante el cálculo de las palabras individuales más repetidas (tokens) en un discurso.

Sin embargo, este tipo de análisis puede carecer de una claridad general de contexto. Un método para tratar de obtener este nivel de matiz se obtiene mediante el análisis de palabras que tradicionalmente podrían considerarse combinadas, llamadas n-grams. Ejemplos de este tipo de palabras serían los United States (Estados Unidos) y Founding Fathers (Padres Fundadores). De hecho, los tokens combinados (palabras múltiples) pueden mejorar la interpretación de la salida. En la práctica se ha indicado que las personas tienden a preferir ver una combinación de términos en lugar de palabras simples

Los N-grams pueden ser combinaciones de dos, tres, cuatro o más palabras. Para esta práctica, se usa una combinación de dos palabras consecutivas.

Como fue descrito en la Sección 1, los Pasos A y B son los mismos procedimientos. El Paso B contiene dos operadores adicionales dentro de la ventana de creación de subprocesos.

Estos dos operadores están específicamente relacionados con el proceso bigrams. Estos son Generate N-Grams (Terms) que tiene un valor máximo de longitud establecido en 2 que indica que está buscando un bigrams (o dos tokens combinados) y Filter Tokens (By Content). Configure los paramteros apropiados para obtener los bigrams. Por lo tanto, este operador garantiza que todos los non-n-grams se excluyan de la salida final, ya que solo el resultado que se desea son palabras que se unen.

Como en el Paso C de la sección previa, escriba los resultados en una hoja de cálculo llamada salidaanalisis.xlsx, y transponga esos resultados a otra hoja de trabajo; el mismo proceso que fue descrito en la sección previa. Estos resultados se transformarán en un informe de visualización usando nubes de palabras en una práctica posterior.

Con esta práctica se finaliza el proceso para conocer los conceptos básicos de recuperación de información mediante SQL, cómo eliminar código HTML, convertir en tokens los documentos y enviarlos a una hoja de cálculo. El objetivo también fue conocer la diferencia en la producción de usar n-grams y no usarlos. También debe experimentar con diferentes tipos de n-grams y también personalizar diccionarios de stopword alternativos para comprender cómo cambia los resultados

Como trabajo final de esta práctica se propone variar los resultados obtenidos aplicando diferentes operadores dentro del operador Process Documents From Data, incluyendo por ejemplo el operador de filtrado por longitud (Filter Tokens (by length)), verificando partes del habla tales como nombres o verbos (operador Filter Tokens (by POS Tags)), o palabras derivadas mediante el operador (Stem (Porter)). Sin embargo, el punto clave no es mostrar cuán sofisticado es su modelo de minería de textos o cuántos operadores han incluido; es únicamente identificar los beneficios de la inclusión de estos operadores para su análisis general.