



UNIVERSIDAD DE GRANADA

TRATAMIENTO INTELIGENTE DE DATOS

Práctica 8

Autor

Antonio José Muriel Gálvez



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Granada, 12 de enero de 2025

Índice

Enunciado	2
Introducción	2
Solución Knime	2
Nodo: Number to String	3
Nodo: Strings to Document	3
Nodo: POS Tagger	3
Nodo: Tagged Document Viewer	3
Nodo: Bag Of Words Creator	4
Nodo: TF y IDF	4
Nodo: Document Vector	5
Nodo: Document Data Extractor	5
Nodo: Column Filter	5
Nodo: Joiner	5
Nodo: Partitioning	6
Nodo: Decision Tree Learner	6
Nodo: Decision Tree Predictor	6
Nodo: Scorer	6

Enunciado

IMDb(<https://www.imdb.com>) es una web de reseñas de películas. Descarga el archivo IMDb.csv con algunos (pocos) comentarios etiquetados como positivos (POS) o negativos (NEG). Desarrolla varios clasificadores para determinar de forma automática si una reseña es positiva o negativa a partir del texto de la misma. Realiza una tarea de clustering sobre la base de datos para determinar en la medida de lo posible en cuántos agrupamientos podemos dividir las reseñas, y qué significado tiene cada agrupamiento.

Introducción

Este proyecto se centra en el desarrollo de un flujo de trabajo en KNIME para clasificar reseñas positivas y negativas utilizando técnicas de clustering. Mediante la aplicación de algoritmos como K-Means, se busca agrupar reseñas similares sin la necesidad de etiquetas previas, lo que permite una categorización automática basada en el contenido textual. Esta aproximación no solo ayuda a identificar patrones latentes en los datos, sino que también proporciona una base para mejorar la comprensión del sentimiento general en grandes volúmenes de texto.

Solución Knime

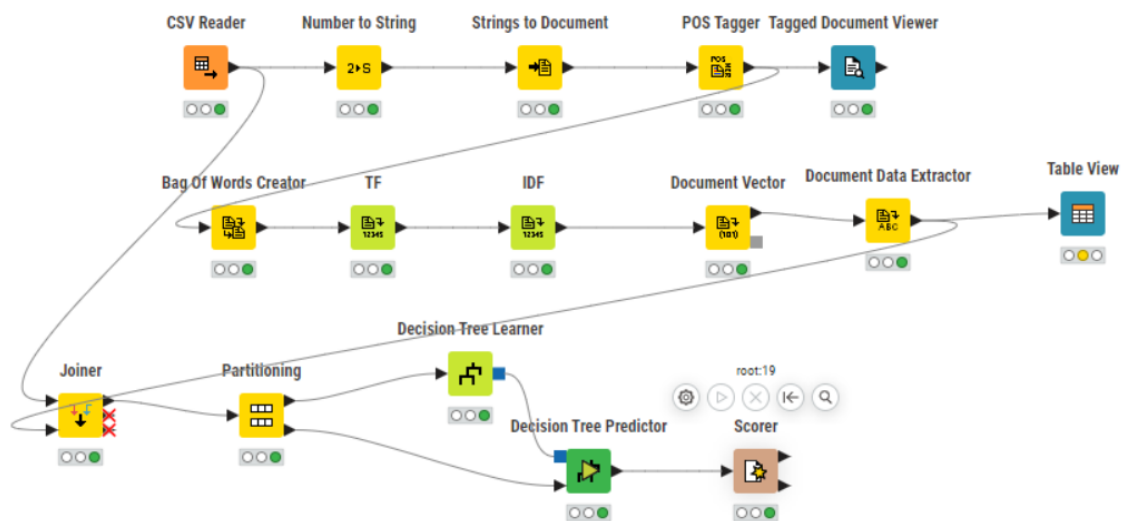


Figura 1: Solución KNIME

- RowID: Identificador único para cada fila.
- Index: Número entero que identifica cada reseña.
- URL: Dirección web relacionada con la reseña.
- Text: Texto de la reseña, que contiene opiniones sobre películas.
- Sentiment: Sentimiento asociado a la reseña (positivo o negativo).

Nodo: Number to String

Este paso se realiza para asegurar que la columna Index, que inicialmente contiene números enteros, sea tratada como texto. Esto puede ser útil para ciertas operaciones posteriores donde el tipo de dato de la columna es importante, como en procesos de concatenación o en nodos que solo admiten datos de tipo String.

Includes: El nodo está configurado para convertir la columna Index (que es de tipo numérico) a tipo String.

Nodo: Strings to Document

El nodo String to Document se utiliza para convertir columnas de texto en documentos que pueden ser procesados por nodos posteriores que trabajan con texto en formato de documento.

Configuración:

- itle: La columna Index se usará como el título de cada documento. Esto ayuda a identificar fácilmente cada documento por un identificador único.
- Full Text: La columna Text se convierte en el contenido principal del documento.
- Use categories from column: Se asignan las categorías o etiquetas desde la columna Sentiment a los documentos. Esto se usa para la clasificación o análisis de sentimientos más adelante.

El nodo genera una nueva columna llamada Document que contiene los datos en formato de documento. Cada documento incluye el título, el texto completo, y las categorías asignadas.

Nodo: POS Tagger

Es un sistema diseñado para analizar comentarios de películas o series, asignándoles una serie de etiquetas estructuradas que facilitan su clasificación y análisis.

Explica su rol en etiquetar partes del habla para análisis más profundo.

Nodo: Tagged Document Viewer

Describe cómo permite revisar el etiquetado de las palabras.

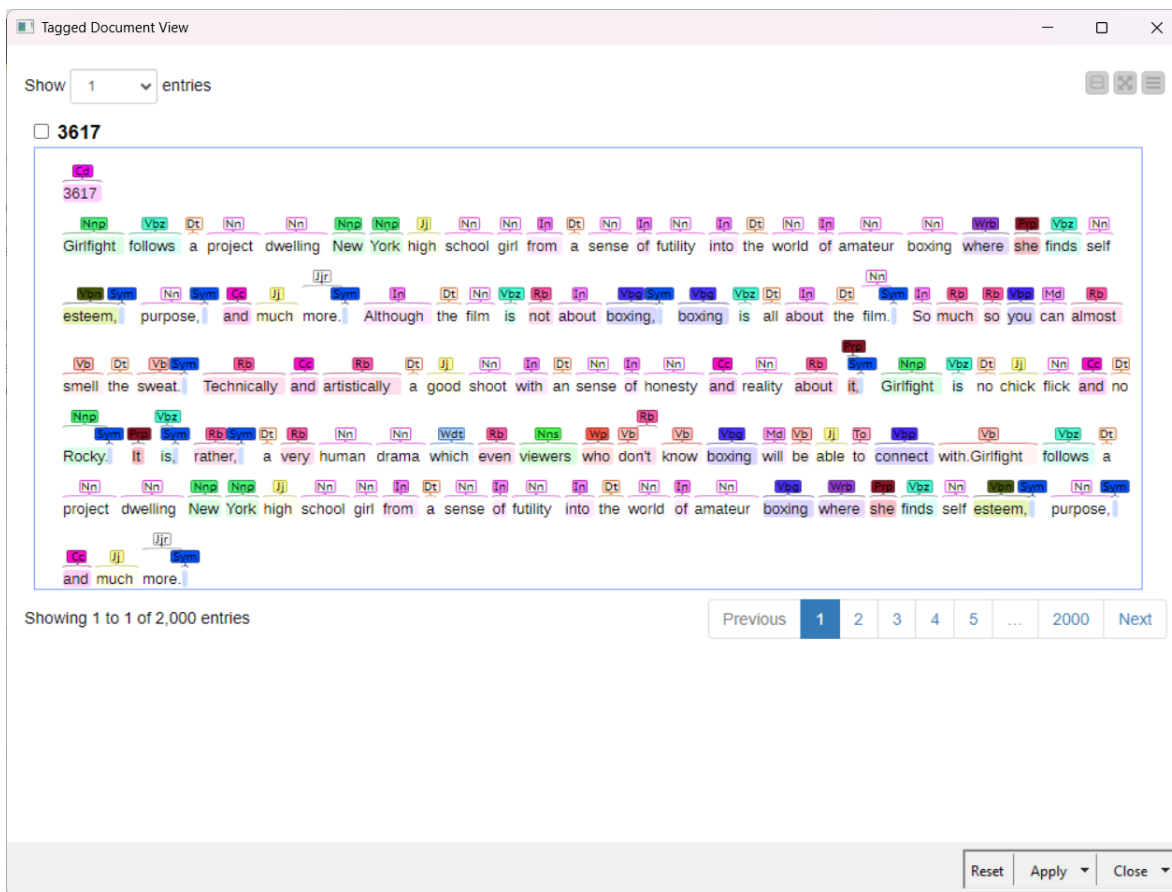


Figura 2: Tagged Document Viewer

Nodo: Bag Of Words Creator

El Bag of Words Creator es un nodo utilizado para descomponer un texto en términos individuales (palabras o tokens) junto con metadatos asociados, como el índice, URL, sentimiento, y más. Este nodo permite representar el texto de manera estructurada para su posterior análisis.

La salida es una tabla con varias columnas que desglosan cada palabra del texto en registros separados. Cada fila en la tabla corresponde a un término individual del texto original. Term: El término individual extraído del texto. Esto incluye el término en sí y puede incluir información adicional, como:

- Tipo gramatical (por ejemplo, sustantivo, verbo).
- Sentimiento asociado.

Nodo: TF y IDF

Los nodos TF (Term Frequency) y IDF (Inverse Document Frequency) son fundamentales en el procesamiento de texto, especialmente en la representación de texto para tareas de minería de datos y machine learning. Se utilizan para calcular la importancia relativa de una palabra dentro de un documento (TF) y en el corpus completo (IDF).

Estos nodos trabajan en conjunto para calcular:

- TF: Frecuencia de un término en un documento individual, lo que indica cuán importante es una palabra en ese documento específico.
- IDF: Frecuencia inversa de documentos, lo que mide la rareza de un término en el conjunto de documentos. Un término que aparece en muchos documentos tendrá un IDF bajo, mientras que un término que aparece en pocos documentos tendrá un IDF alto.

Nodo: Document Vector

Transforma los documentos textuales en vectores numéricos, permitiendo que los documentos sean utilizados en algoritmos de aprendizaje automático y análisis cuantitativo. Esto es esencial para convertir el texto no estructurado en una forma que los modelos puedan entender.

Bitvector (activado): Esta opción genera una representación binaria de los términos en los documentos. Si un término está presente en un documento, se asigna un 1, y si no está presente, se asigna un 0.

La salida es una matriz de documentos donde cada documento está representado como un vector de características numéricas.

Nodo: Document Data Extractor

Se utiliza para extraer datos estructurados de documentos textuales. Este nodo permite extraer partes específicas del documento, como el texto completo, términos, o cualquier metadato adicional que se haya generado durante el procesamiento de texto.

Configuración del Nodo:

- Document column: Document. La columna que contiene los documentos textuales procesados. Data extractors: Text. Extrae el contenido textual completo de cada documento.
- Separate terms by whitespaces: Activado. Esto indica que los términos deben ser separados por espacios en blanco, lo que facilita su análisis posterior.

Nodo: Column Filter

El nodo Column Filter se utiliza para seleccionar, eliminar o reordenar columnas de una tabla de datos. En este contexto, se utiliza para asegurarse de que sólo las columnas necesarias para el siguiente paso (específicamente la columna Sentiment) se mantengan.

Incluir Sentiment, la columna que contiene las etiquetas de clasificación, Positivo o Negativo. Excluir todas las demás columnas que no son necesarias para la operación de unión posterior.

Nodo: Joiner

Se utiliza para combinar dos tablas de datos basadas en una o más columnas clave comunes. En este flujo de trabajo, se utiliza para volver a añadir la columna Sentiment a la tabla de características generada por el nodo Document Data Extractor.

Nodo: Partitioning

Divide un conjunto de datos en dos subconjuntos. Este proceso es crucial para tareas de modelado y evaluación, permitiendo separar los datos en conjuntos de entrenamiento y prueba.

La partición se hará en función de un porcentaje relativo del total de datos:

- Porcentaje para el Primer Conjunto: El primer conjunto, que se utiliza como conjunto de entrenamiento recibirá el 80 % de los datos totales.
- Porcentaje para el Segundo Conjunto: El segundo conjunto, utilizado como conjunto de prueba, recibirá el 20 % restante de los datos.

Nodo: Decision Tree Learner

Usado para construir un modelo de árbol de decisión basado en datos de entrenamiento. Este modelo se usa para predecir una variable objetivo categórica, Sentiment en este caso, a partir de un conjunto de características.

En la configuración específica que la columna Sentiment es la variable objetivo que el árbol de decisión intentará predecir y contiene valores categóricos como POS (positivo) o NEG (negativo), que el modelo clasificará.

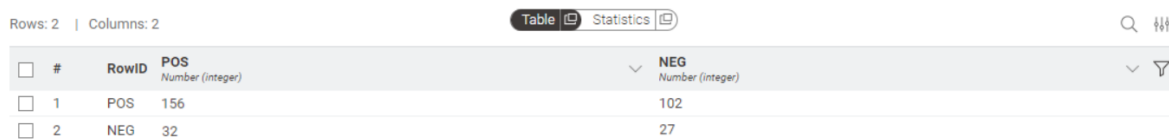
Nodo: Decision Tree Predictor

El nodo Decision Tree Predictor en KNIME se utiliza para aplicar el modelo de árbol de decisión entrenado a un conjunto de datos nuevos, normalmente el conjunto de prueba, para predecir la clase de cada instancia.

Nodo: Scorer

evaluar el rendimiento de un modelo de clasificación comparando las etiquetas reales con las predicciones hechas por el modelo.

- First Column: Sentiment, columna real con las etiquetas de clase.
- Second Column: Prediction(Sentiment), columna con las predicciones del modelo.



Rows: 2 | Columns: 2

	#	RowID	POS Number (integer)	NEG Number (integer)
<input type="checkbox"/>	1	POS	156	102
<input type="checkbox"/>	2	NEG	32	27

Figura 3: Scorer