UNIVERSIDAD DE GRANADA

decsai.ugr.es

Minería de Medios Sociales

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Bloque II: Minería de Texto y de la Web

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Sesión 5:
Minería de Opiniones y Sentimientos con KNIME

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

Análisis de Sentimientos en KNIME

KNIME (KoNstanz Information MinEr)

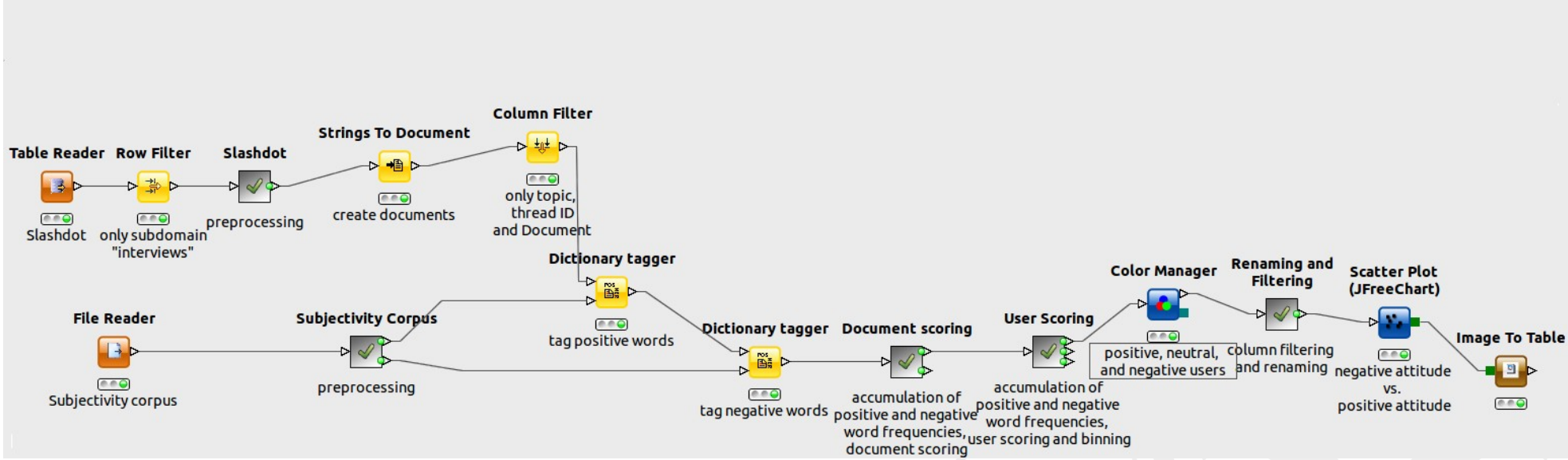
Workflows:

https://www.knime.org/white-papers

Social Media →
Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining (2012) →
Social Media Sentiment Analysis Example Workflow

https://www.knime.org/blog/sentiment-analysis

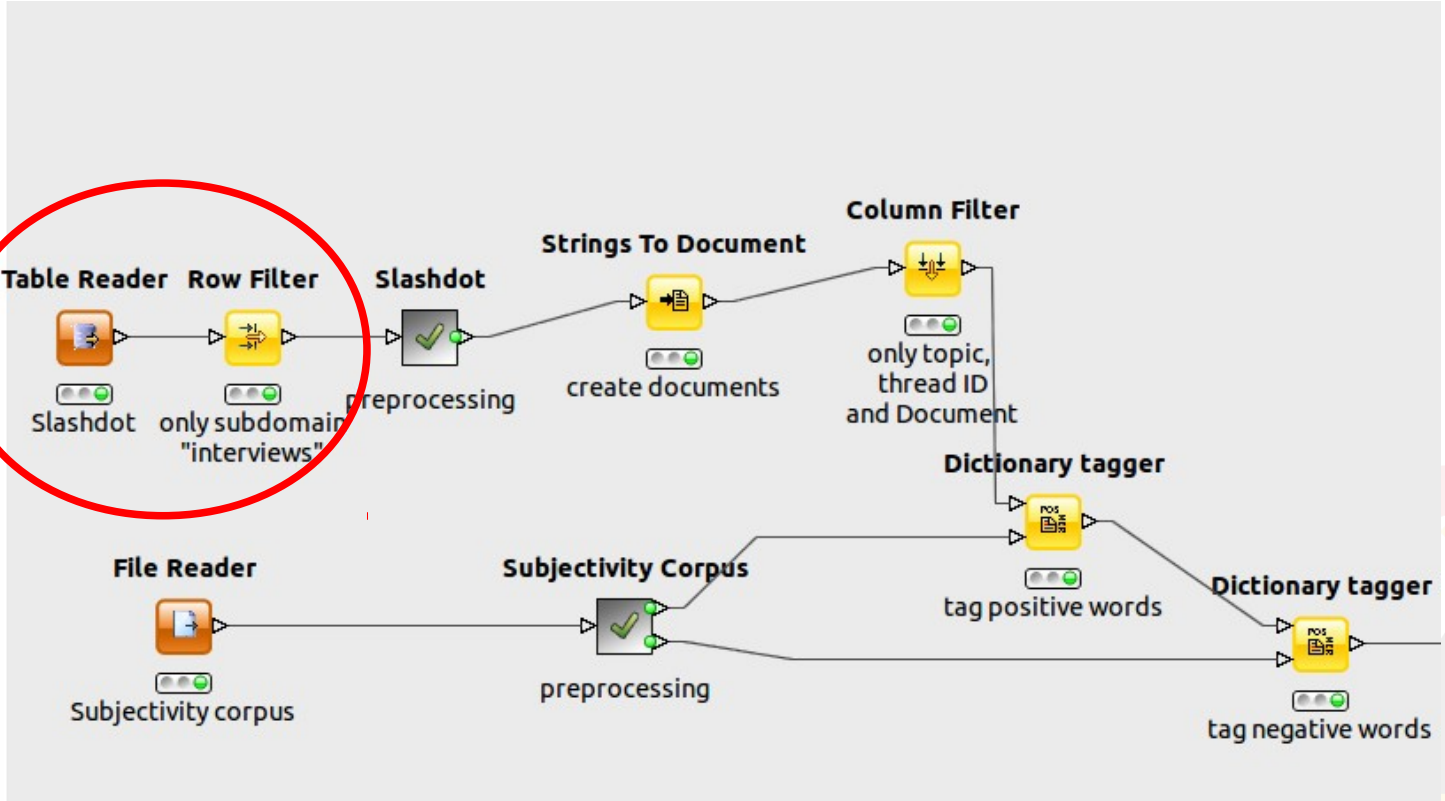
Social Media Sentiment Analysis



A partir de post de Slashdot, y usando el MPQA Subjectivity Lexicon (http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/) se analiza la actitud de usuarios identificados en el sistema como positiva, neutral o negativa.

Los datos obtenidos se muestran en una gráfica donde podemos observar los usuarios analizados y su actitud usando un código de colores.

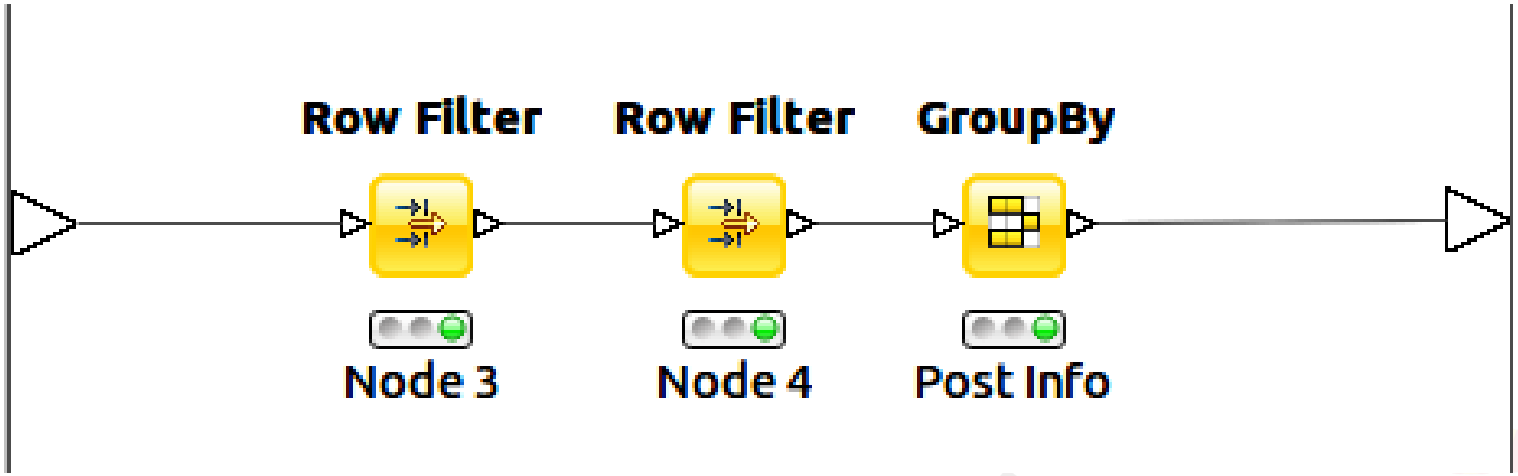
Obtención de Datos



Los datos de Slashdot se cargan a partir de una tabla en disco usando TableReader.

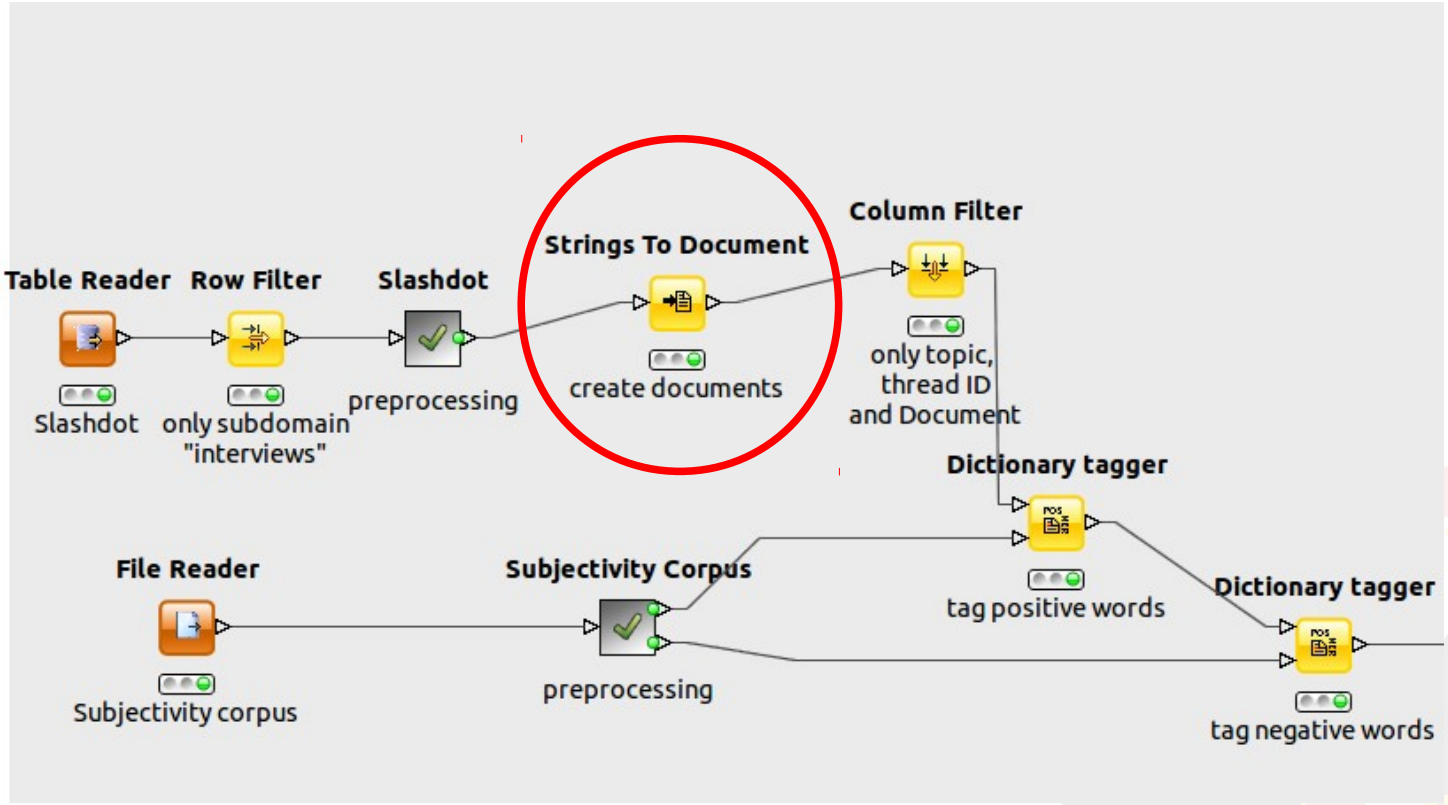
Los datos están categorizados, en este caso se van a analizar los datos de la categoría “Interviews”

Selección de Usuarios



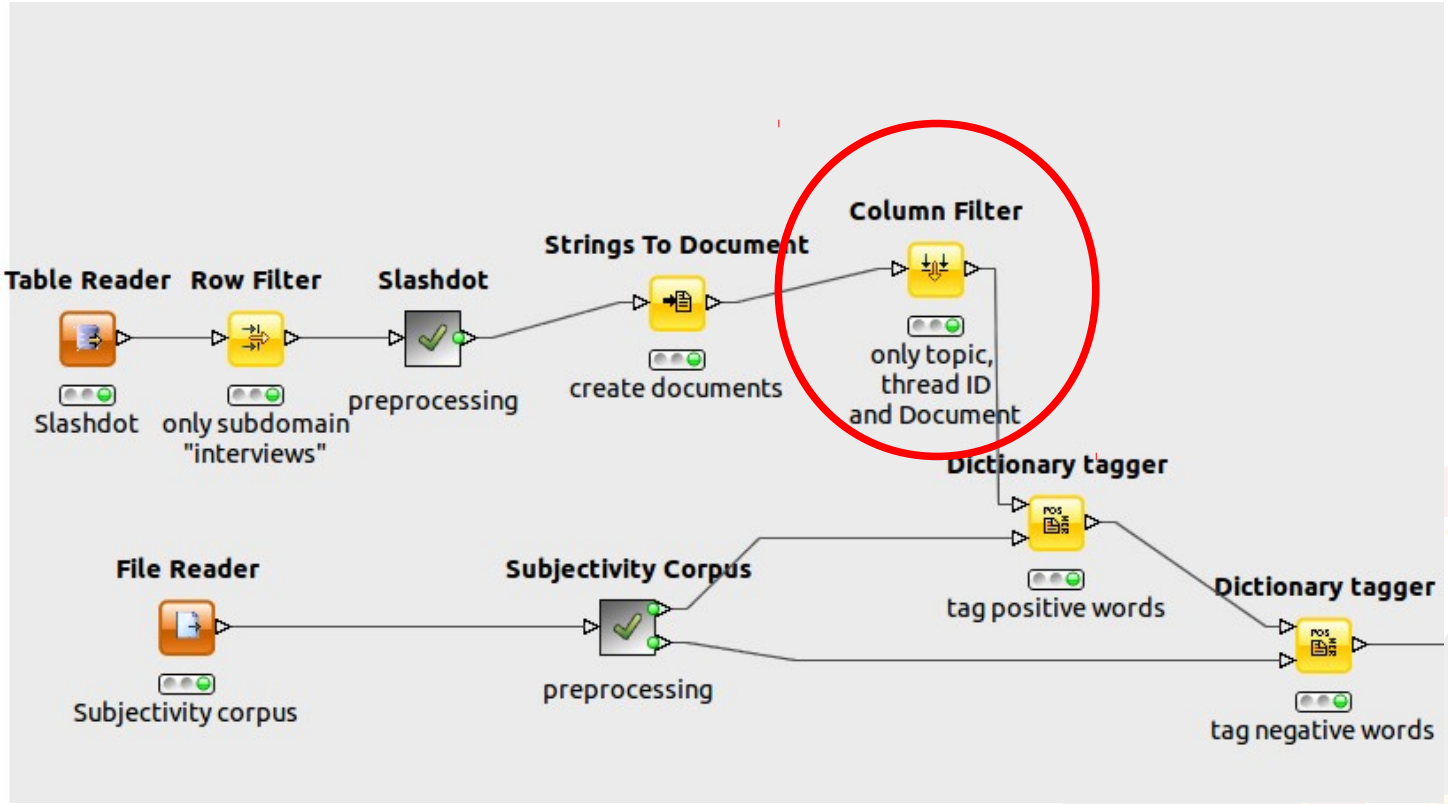
- Seleccionamos a los usuarios autentificados en el sistema.
- Eliminamos los usuarios anónimos.
- Agrupamos todos sus comentarios para un determinado post

Generamos los Documentos



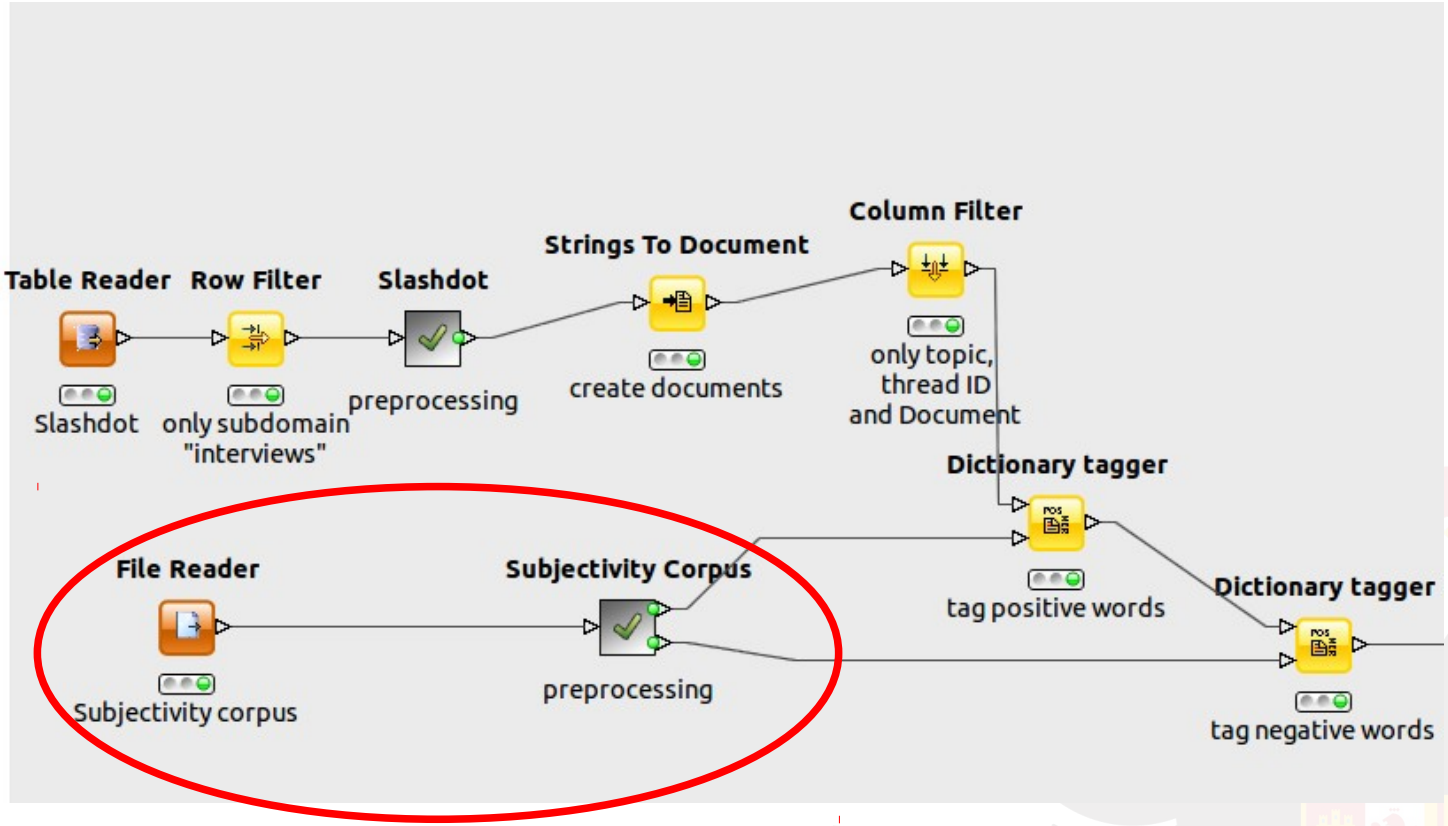
Generamos una columna adicional con el documento creado a partir de los datos.

Generamos los Documentos



Nos quedamos con las columnas que contienen el identificador de la discusión, el tema y el Documento.

Carga del Corpus de Sentimiento



Leemos el corpus a partir de un fichero y lo preprocesamos.

Obtenemos la Tabla del Fichero

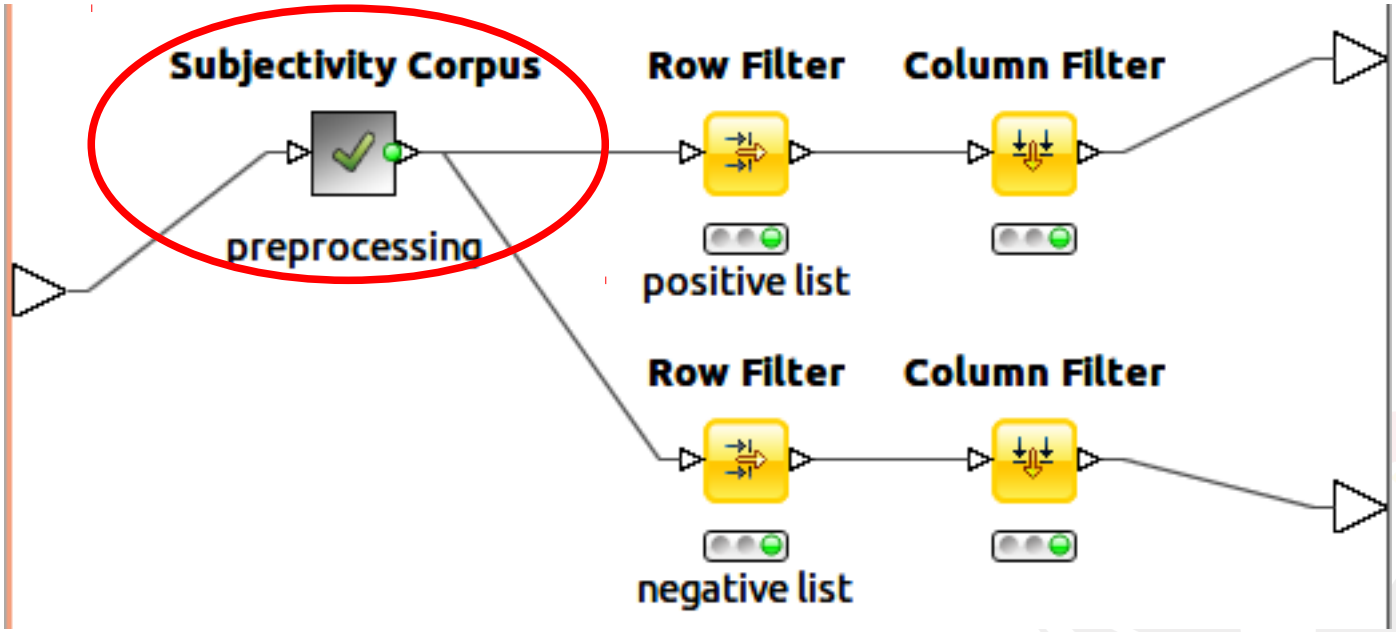
File Table - 0:25 - File Reader(Subjectivity corpus)

File

Table "subjclueslen1-HLTEMNLP05.tff" - Rows: 8221 | Spec - Columns: 6 | Properties | Flow Variables

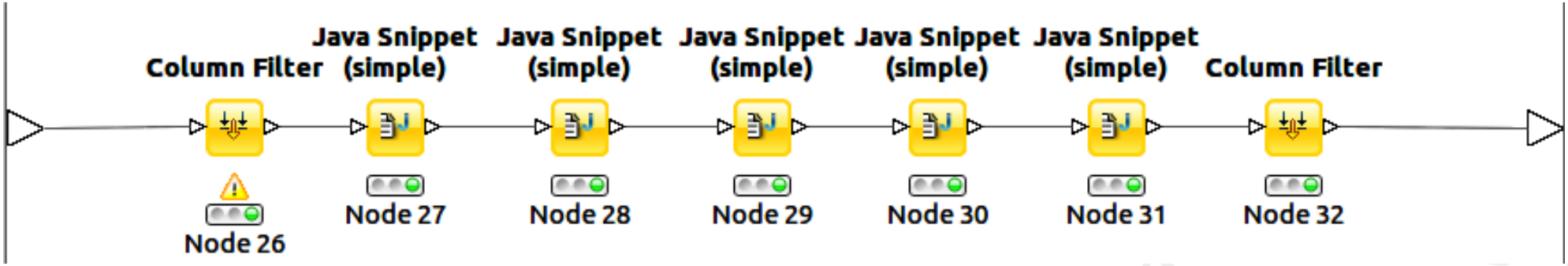
Row ID	Col0	Col1	Col2	Col3	Col4	Col5
Row0	type=weaks...	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
Row1	type=weaks...	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
Row2	type=weaks...	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
Row3	type=strong...	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
Row4	type=strong...	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
Row5	type=strong...	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
Row6	type=weaks...	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
Row7	type=weaks...	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
Row8	type=strong...	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
Row9	type=strong...	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative
Row10	type=strong...	len=1	word1=abhor	pos1=anypos	stemmed1=y	priorpolarity=negative
Row11	type=strong...	len=1	word1=abhor	pos1=verb	stemmed1=y	priorpolarity=negative
Row12	type=strong...	len=1	word1=abhorred	pos1=adj	stemmed1=n	priorpolarity=negative
Row13	type=strong...	len=1	word1=abhorrence	pos1=noun	stemmed1=n	priorpolarity=negative
Row14	type=strong...	len=1	word1=abhorrent	pos1=adj	stemmed1=n	priorpolarity=negative
Row15	type=strong...	len=1	word1=abhorrently	pos1=anypos	stemmed1=n	priorpolarity=negative
Row16	type=strong...	len=1	word1=abhors	pos1=adj	stemmed1=n	priorpolarity=negative

Preprocesamos el Corpus



Vamos a adaptar el corpus para poder usarlo.

Adaptar el Corpus



Preprocesamos el corpus:
Por cada una de las columnas originales creamos una nueva en la que vamos a eliminar el nombre del campo que está dentro del texto.

Filtered table - 0:192:187:32 - Column Filter

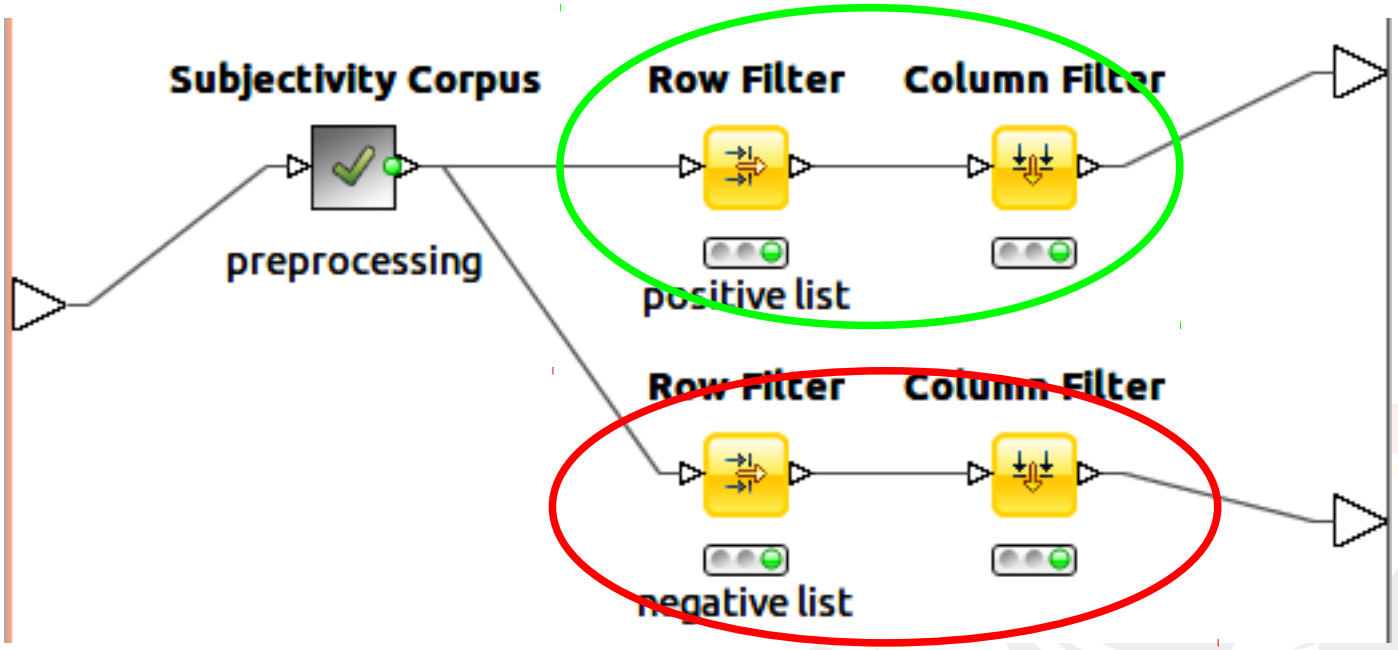
File

Properties | Flow Variables

Table "default" - Rows: 8221 | Spec - Columns: 5

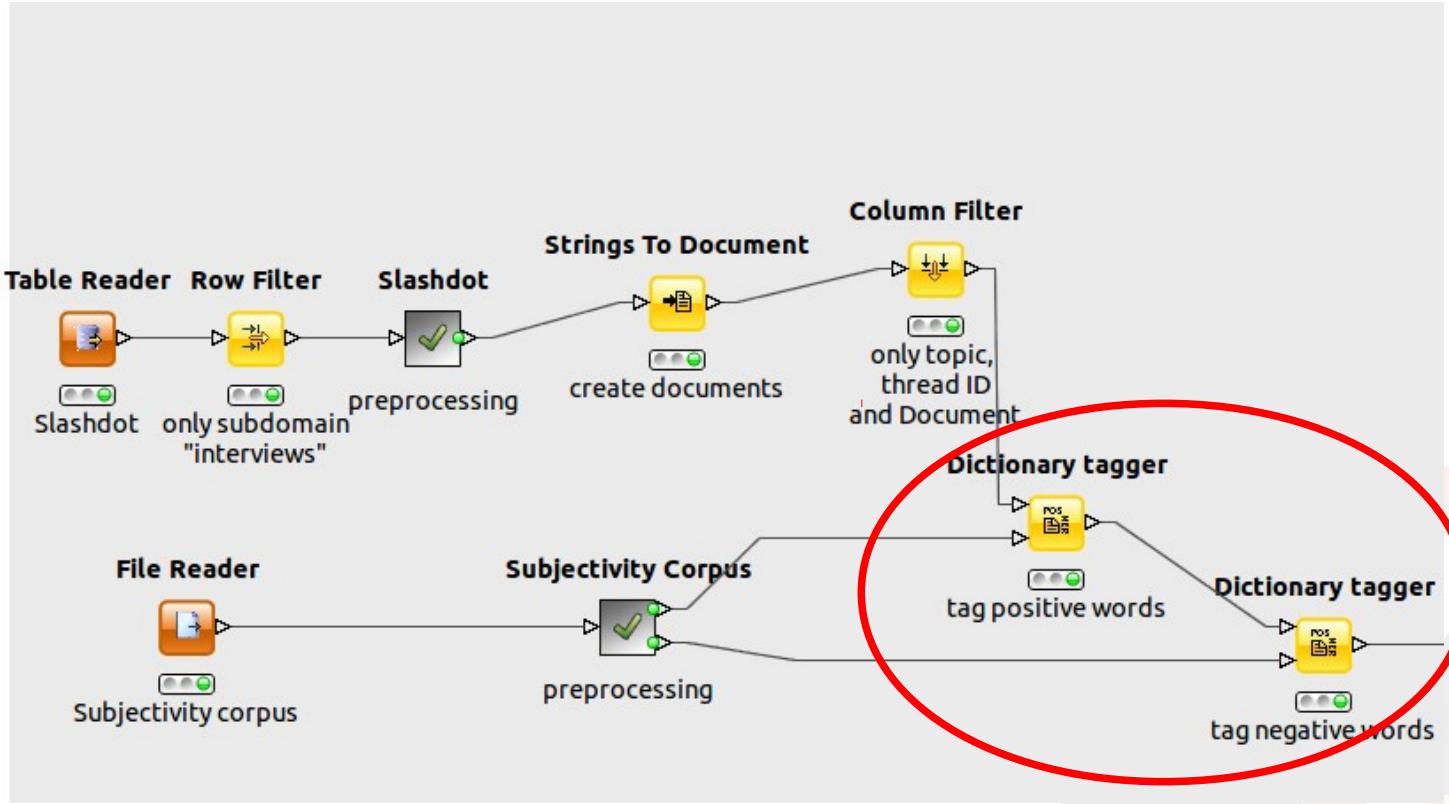
Row ID	S Type	S Word	S Pos	S Stem...	S Polartity
Row0	weaksubj	abandoned	adj	n	negative
Row1	weaksubj	abandon...	noun	n	negative
Row2	weaksubj	abandon	verb	y	negative
Row3	strongsubj	abase	verb	y	negative

Seleccionamos las Columnas Útiles



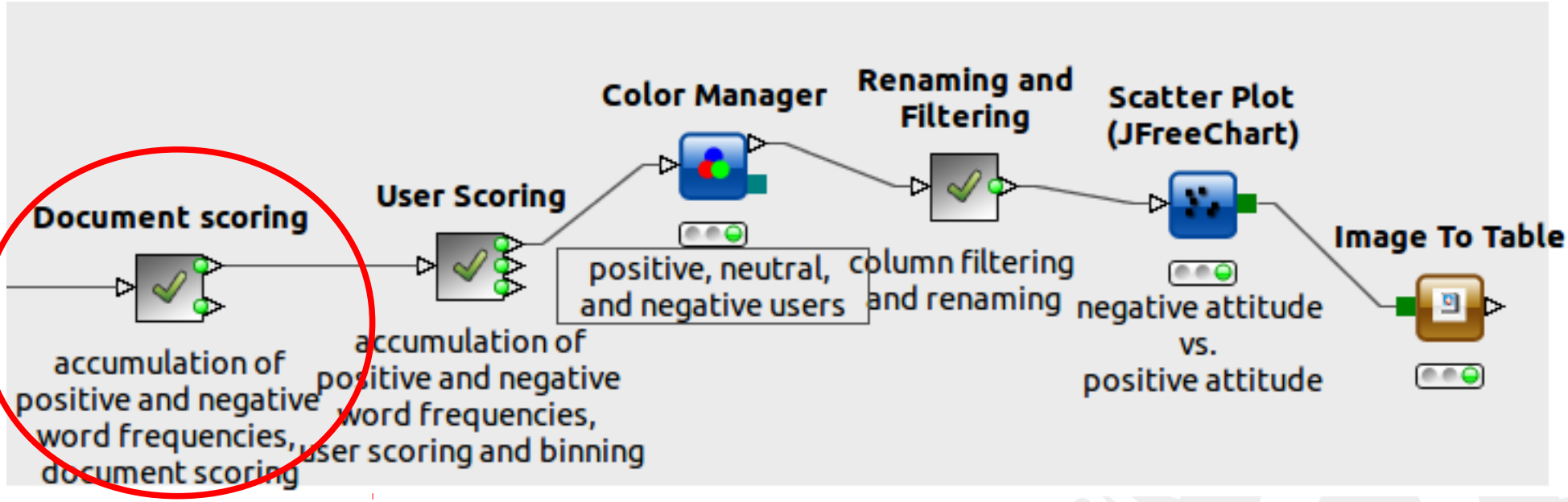
Nos quedamos con la palabras de polaridad positiva y con las de negativa

Etiquetado de los Datos



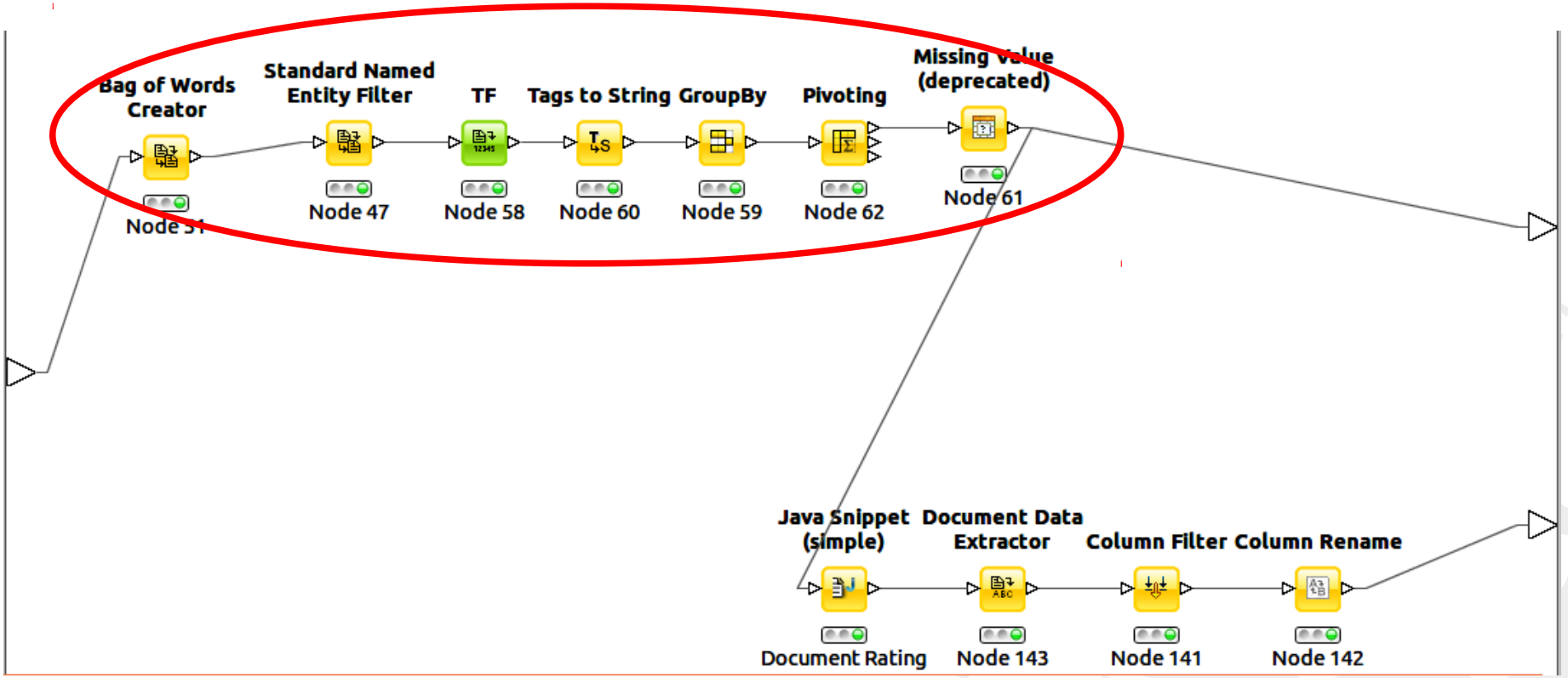
Se etiquetan las apariciones de palabras positivas como PERSON y de negativas como MONEY

Cálculo de Pesos y Generación de Gráficas



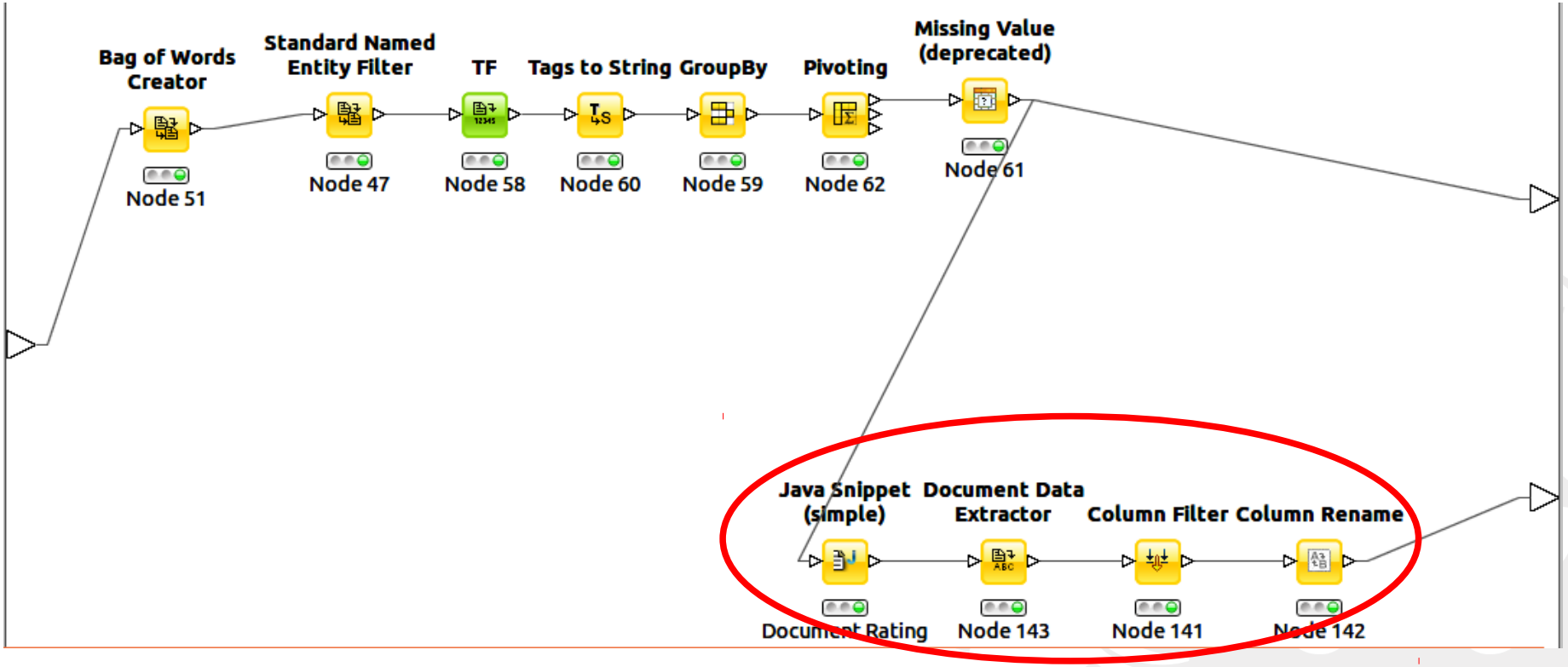
Se calcula el número de palabras positivas y negativas que aparecen en el documento.

Frecuencia de las Palabras de Sentimiento



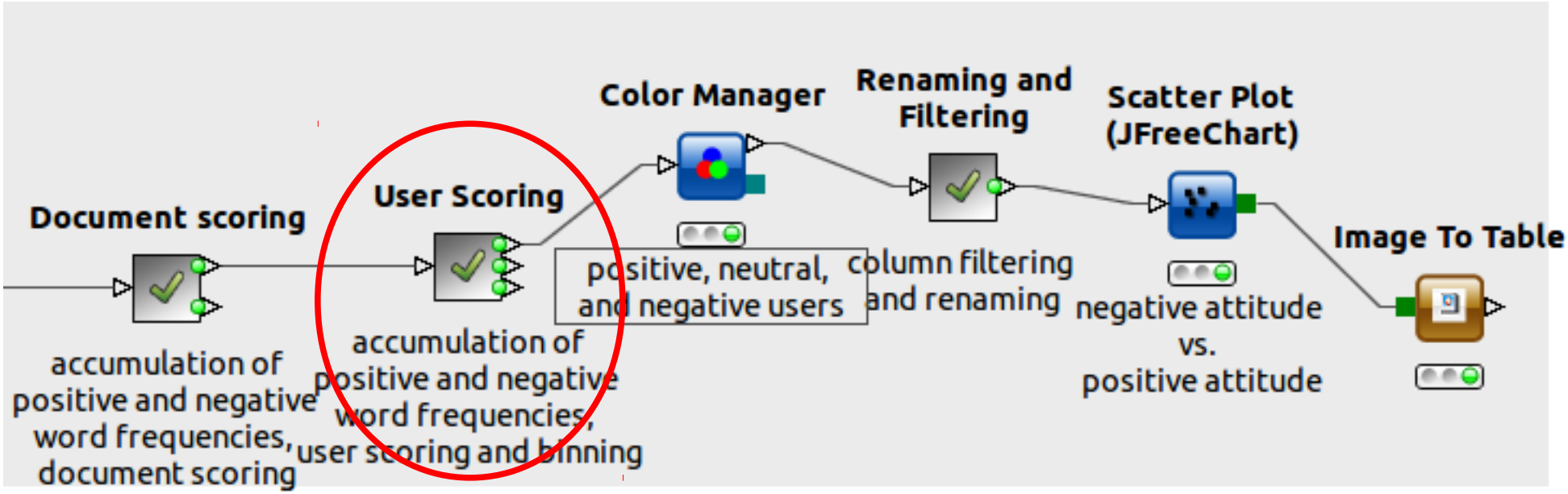
Se genera la representación en forma de bolsa de palabras, se seleccionan únicamente aquellos elementos que están etiquetados como PERSON y MONEY y se cuentan las frecuencias. Se convierten las etiquetas a texto y se agrupan según el valor de las etiquetas, se agregan y suman los valores. Los valores que no aparecen, se consideran 0.

Orientación del Sentimiento de Documentos



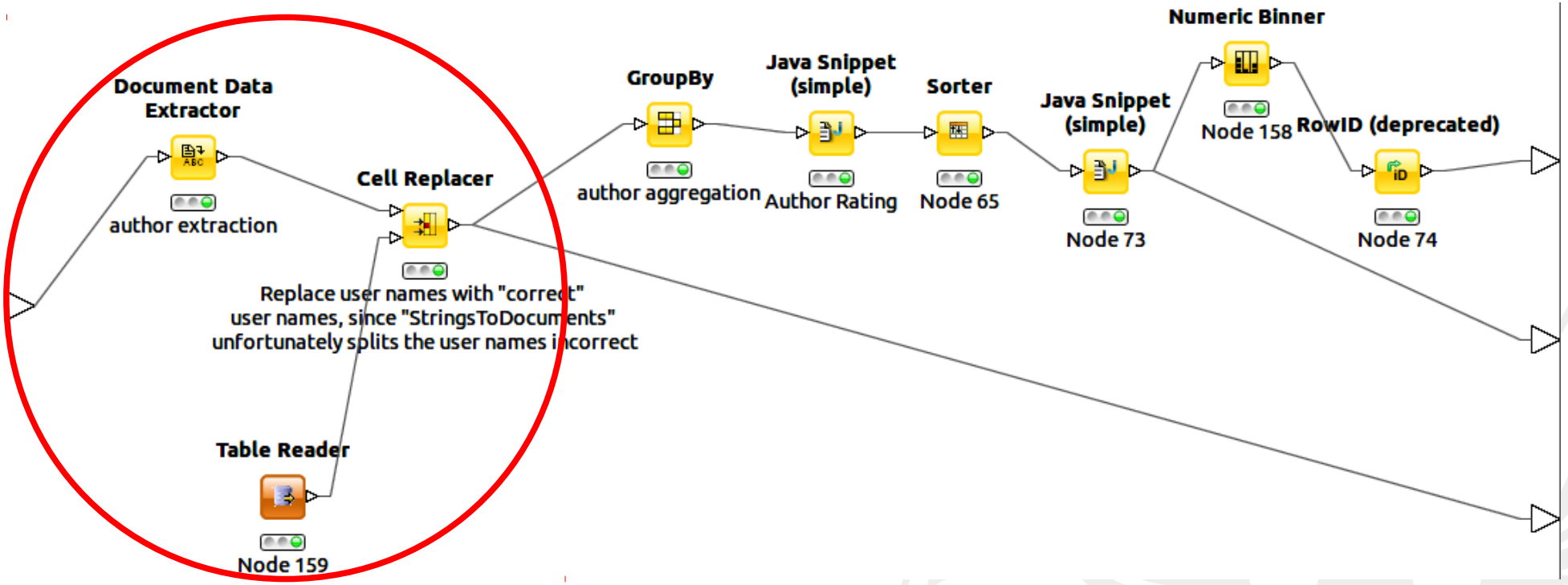
Se calcula la diferencia entre las etiquetas positivas y negativas, obtenemos el título, el valor para etiquetas positivas, para negativas y para la diferencia, se renombran las columnas para facilitar la comprensión y manejo de los datos.

Pesos de los Usuarios



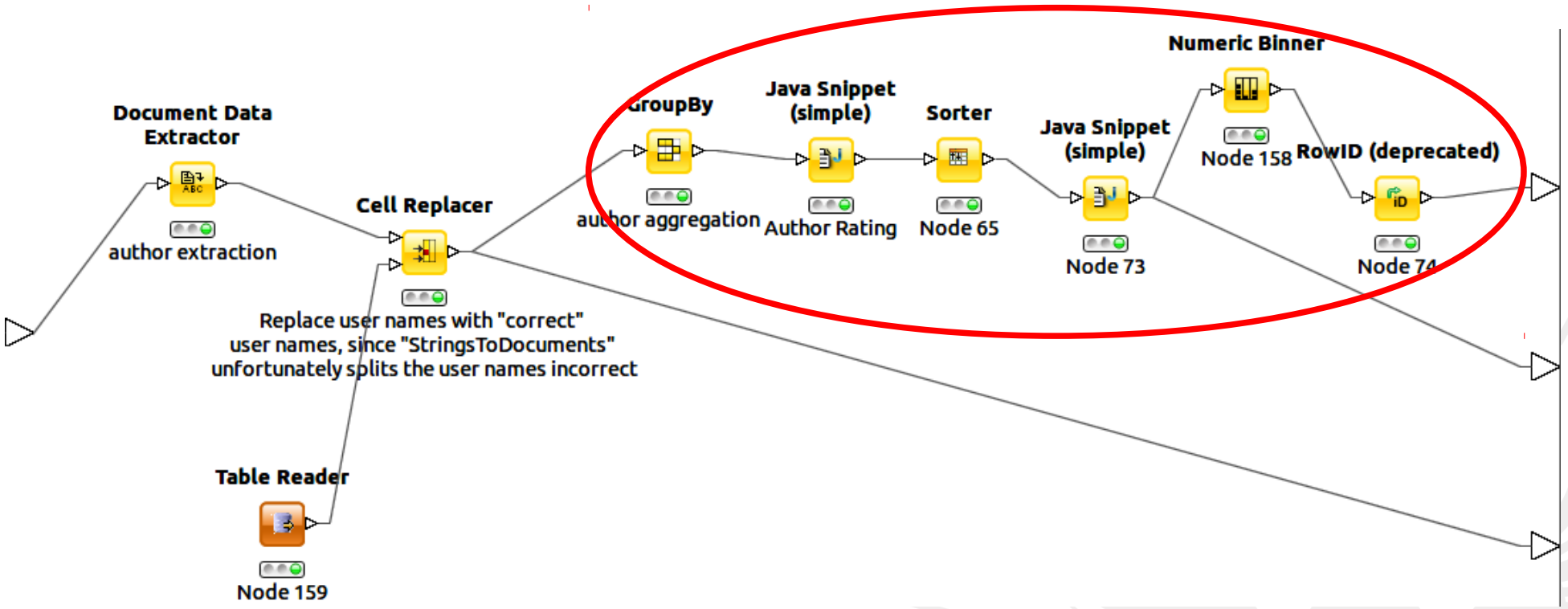
Calculamos los pesos para cada usuario.

Procesamiento de Usuarios



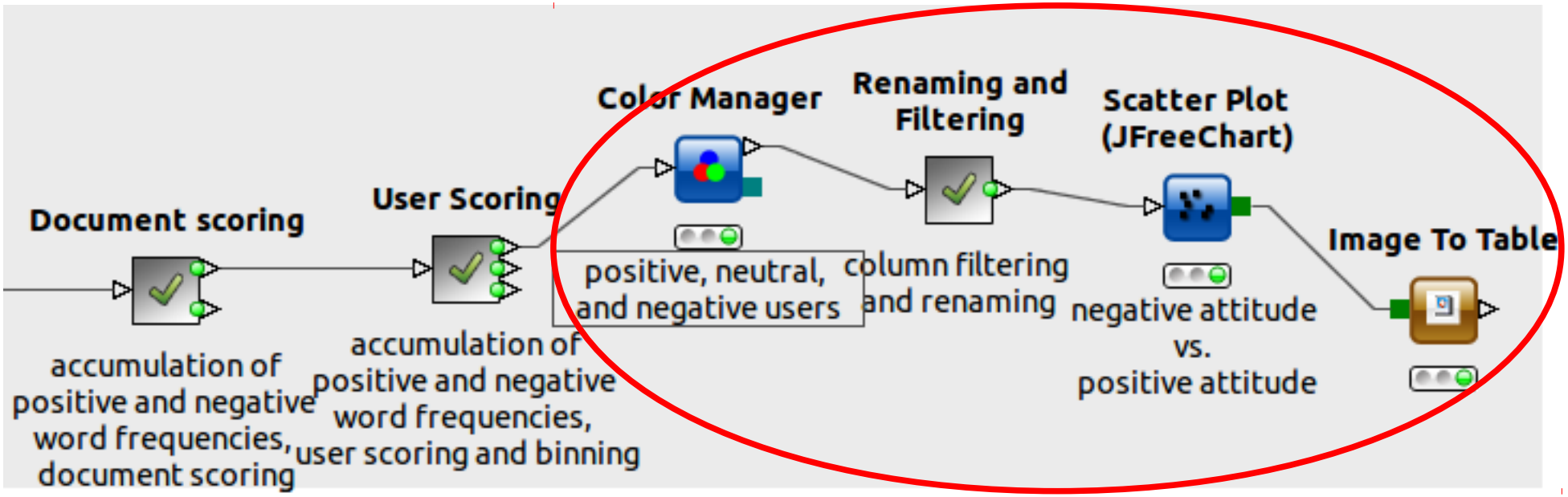
Leemos una tabla con los nombres de los usuarios de disco, sustituimos los nombre actuales por los leídos de la tabla, ya que los originales se han visto afectados en el procesamiento.

Orientación del Sentimiento



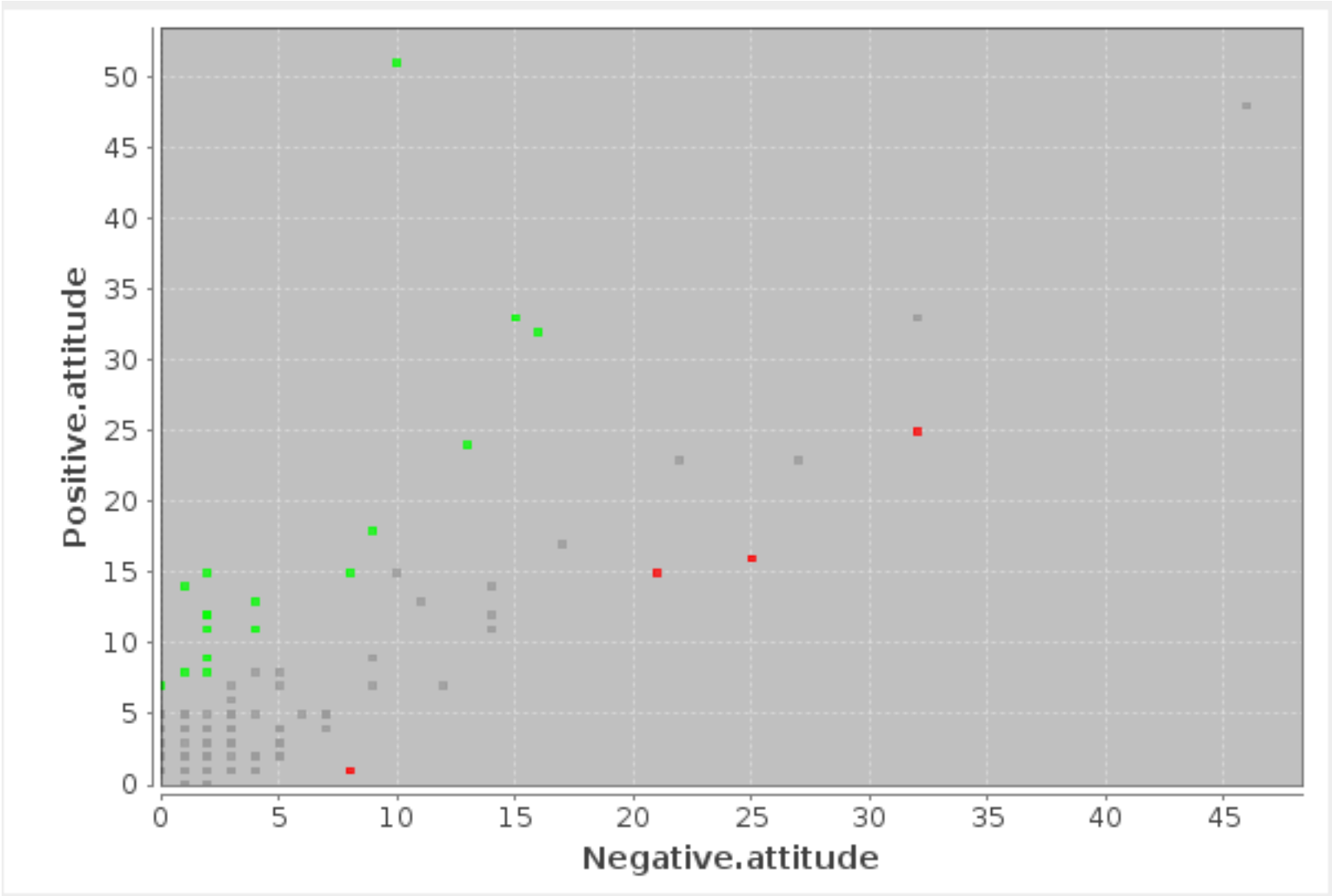
Se agrupa por autor y se calcula la diferencia entre orientaciones en los sentimientos. Se realiza la ordenación y posteriormente se reducen los valores a positivo o negativo según sea el valor numérico. Se asocian rangos de valores a etiquetas positivo, negativo, neutral.

Generación de Gráficas



Se asocia a cada orientación del sentimiento para un usuario un color, se renombran las columnas para facilitar la comprensión y se generan los gráficos.

Gráfica Resultante



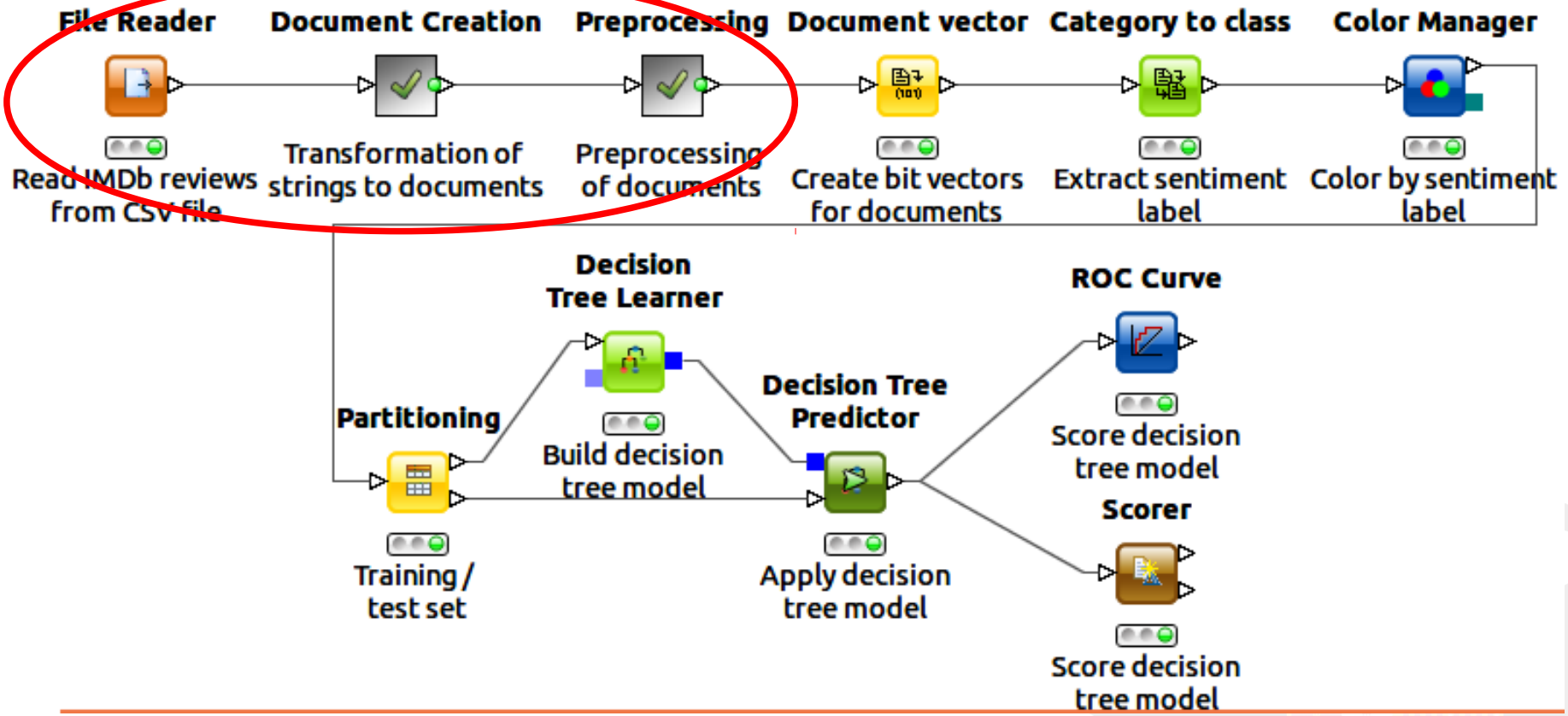
Clasificación de Sentimientos

A partir de un conjunto de datos obtenidos de IMDb previamente clasificados (1000 positivos, 1000 negativos), trataremos de asignar la etiqueta correspondiente a cada texto.

Preprocesaremos el texto y lo clasificaremos utilizando un conjunto de entrenamiento del 70% y un conjunto de prueba del 30%.

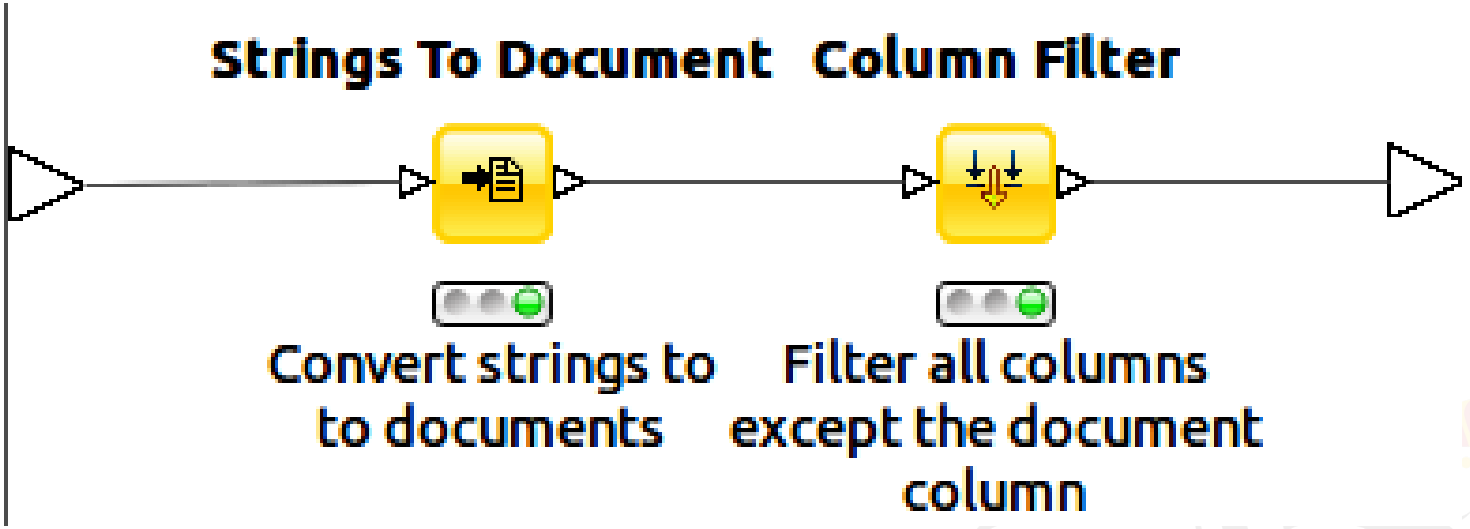
Mostraremos una curva ROC (Receiver Operating Characteristic) para visualizar la precisión de la clasificación.

Clasificación de Sentimientos



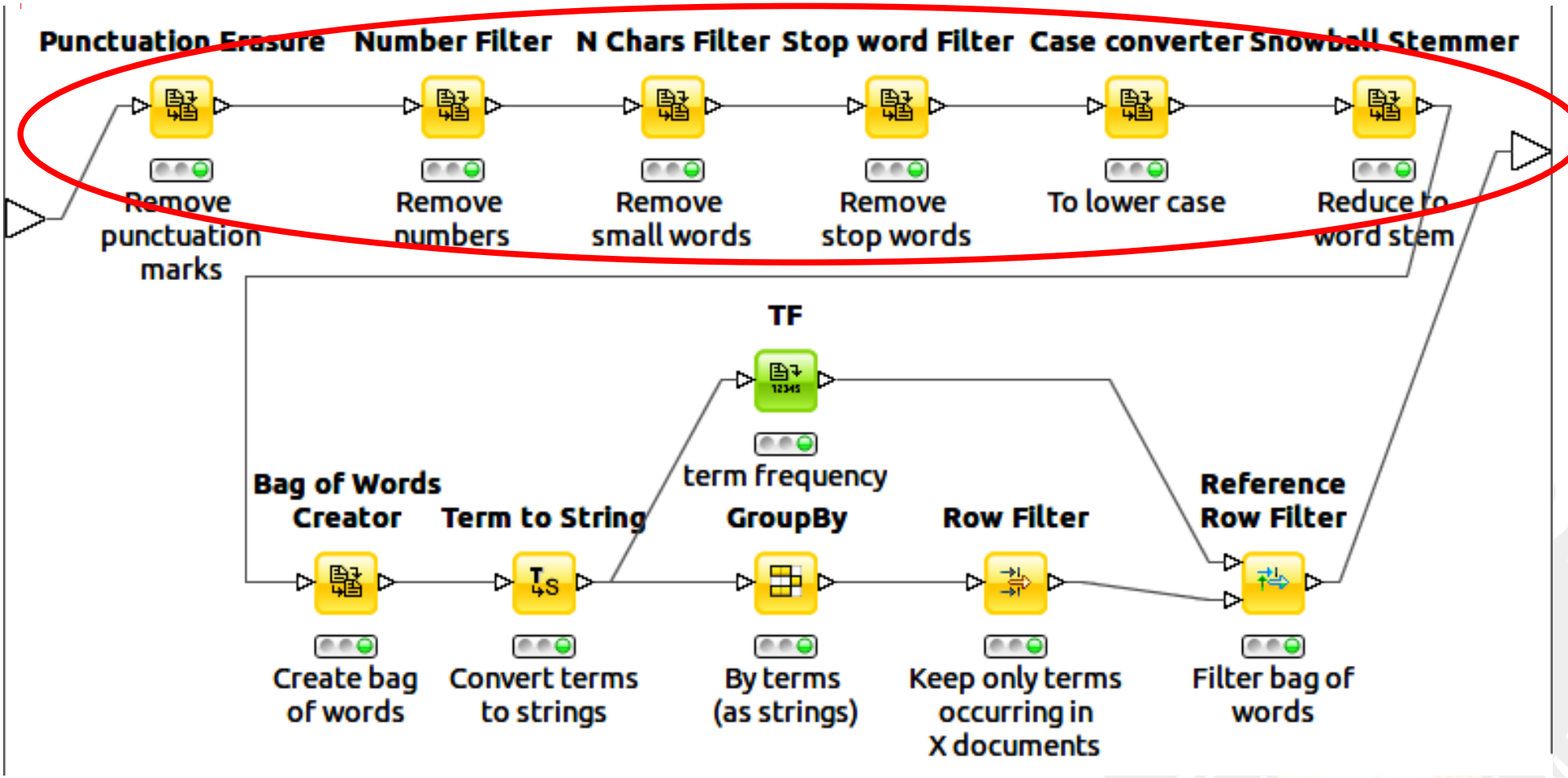
Leemos datos que contienen comentarios sobre películas de IMDb de un fichero CSV, creamos los documentos y los preprocesamos.

Creación de Documentos



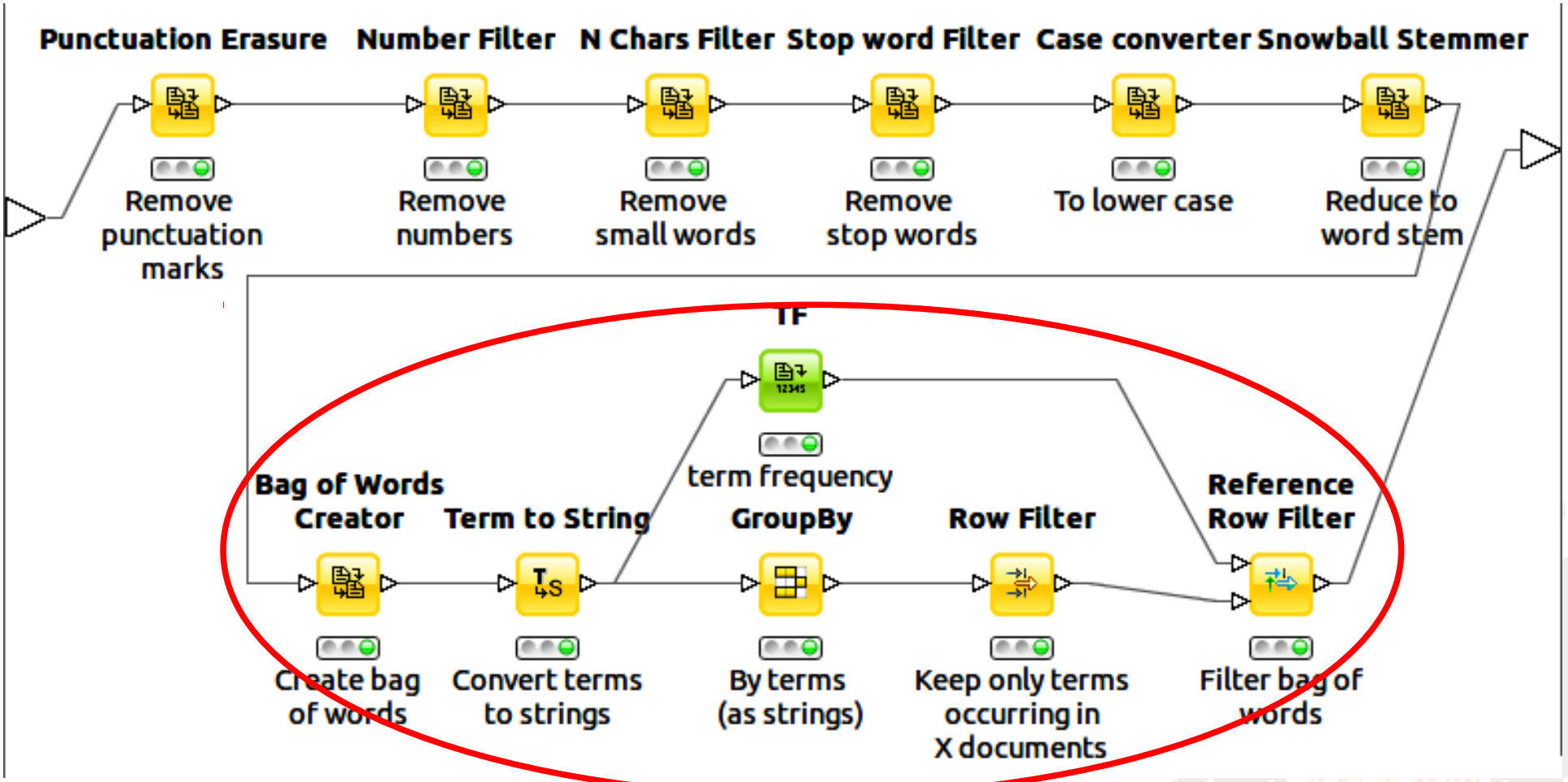
Generamos el documento a partir de los datos de la tabla y nos quedamos únicamente con la columna que contiene los documentos.

Preprocesamiento



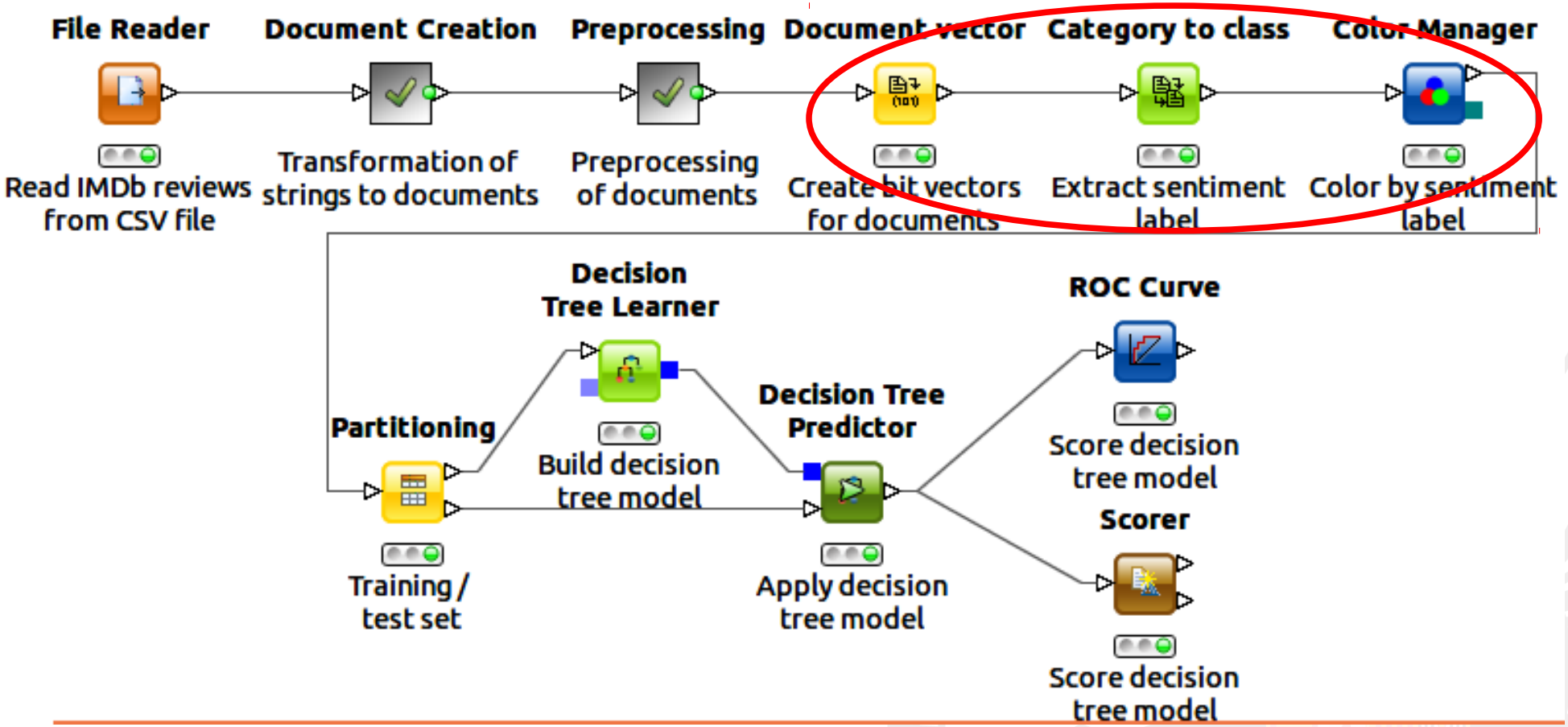
Aplicamos tokenización, eliminamos números, eliminamos palabras de menos de 2 caracteres, eliminamos stop-words, ponemos todo en minúscula y aplicamos lematización.

Preprocesamiento



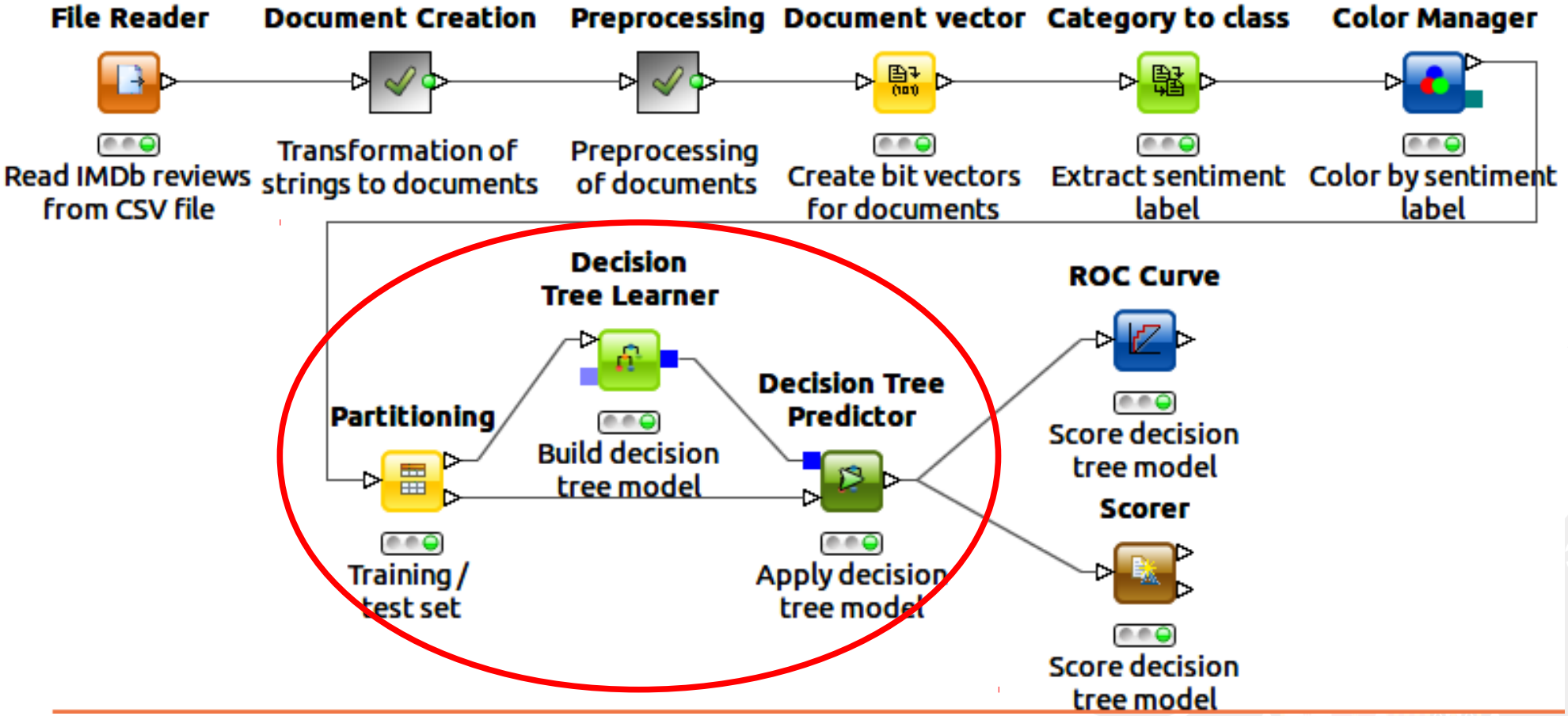
Generamos una bolsa de palabras, convertimos los términos en cadenas. Dividimos el flujo, por una parte contabilizamos las frecuencias, y por otra agregamos los términos y nos quedamos con aquellos que aparezcan en al menos 20 documentos. Finalmente mezclamos ambos flujos y obtenemos las frecuencias de los términos que aparecen en más de 20 documentos.

Identificación de Clases



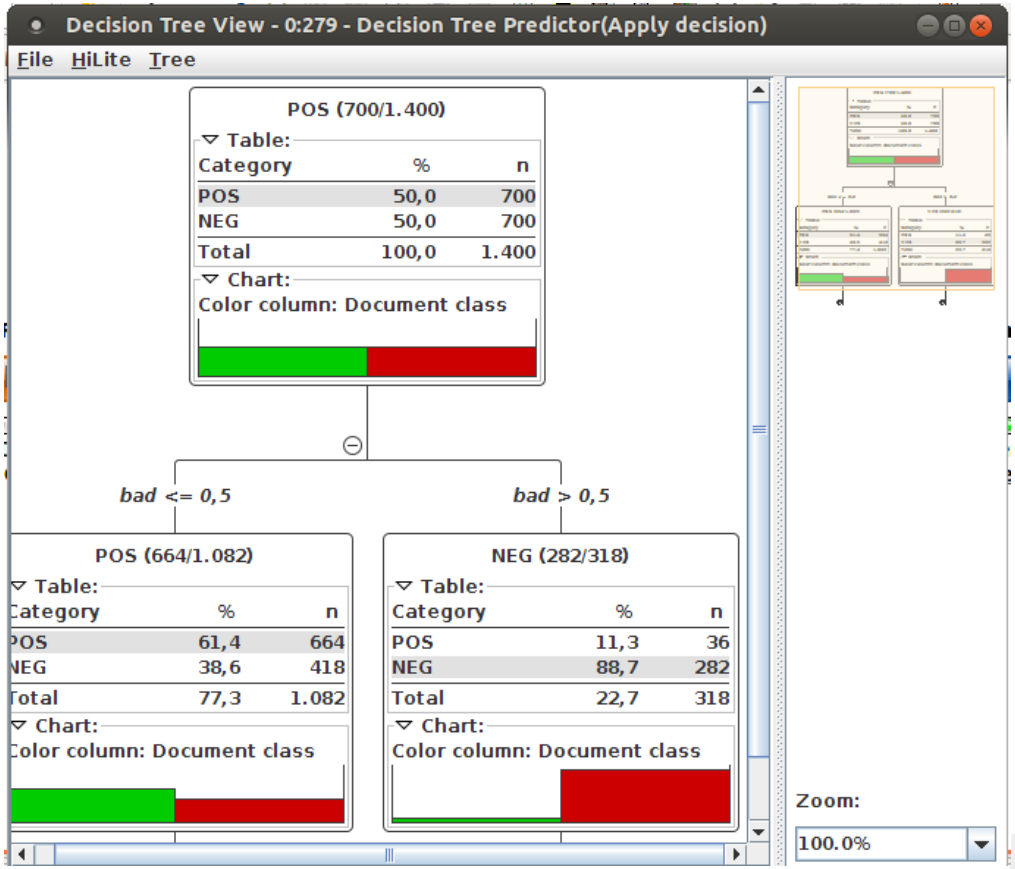
Pasamos la categoría del documento a una clase, que se añade como una nueva columna. En base a esa clase se asigna un color, verde para positivo (POS), rojo para negativo (NEG).

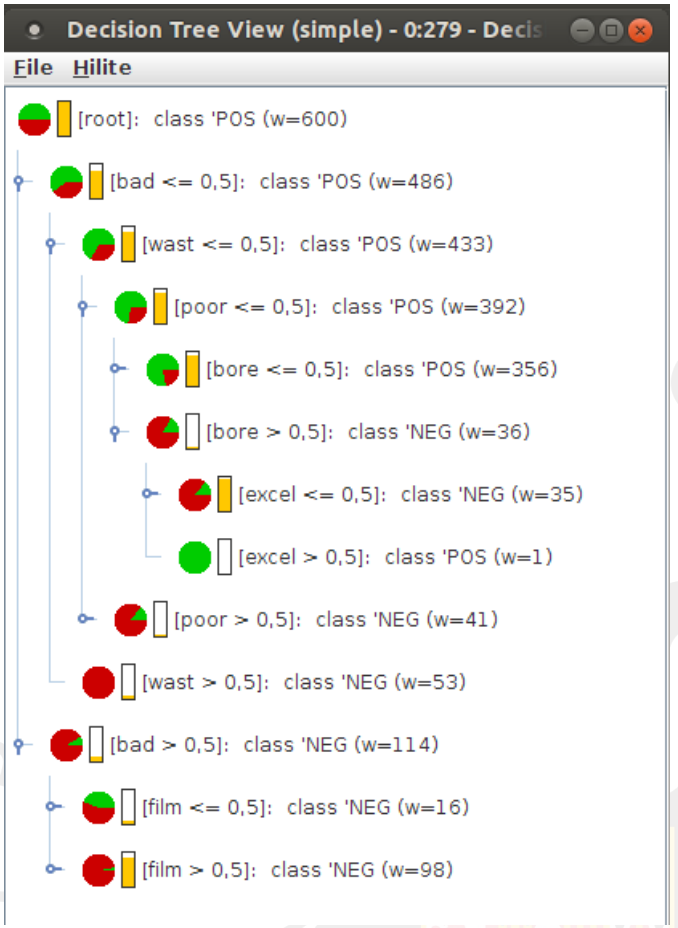
Clasificación



Partimos el conjunto de datos, y pasamos el 70% al clasificador como conjunto de entrenamiento, y el 30% restante se usa para prueba, y se realiza la clasificación.

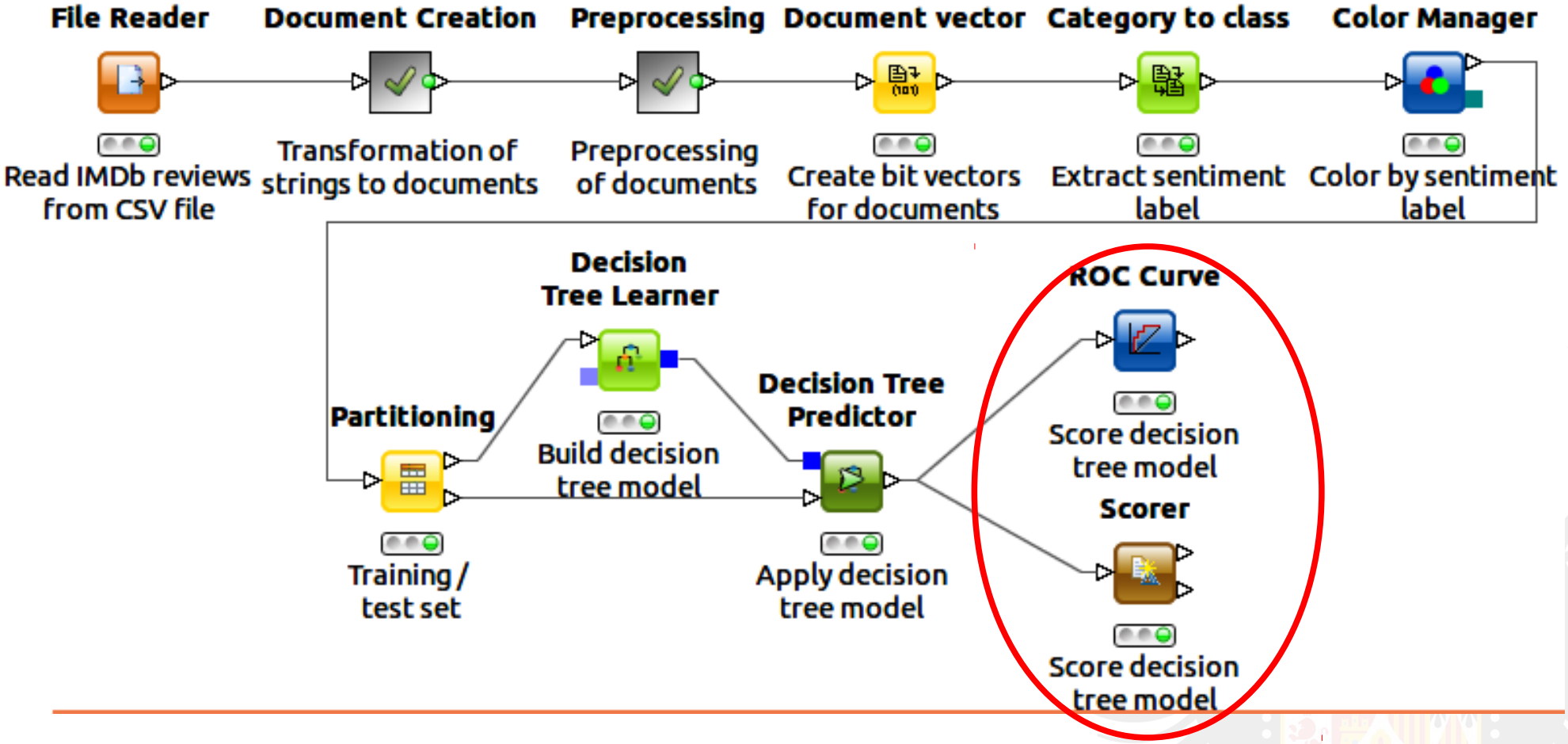
Árbol de Decisión





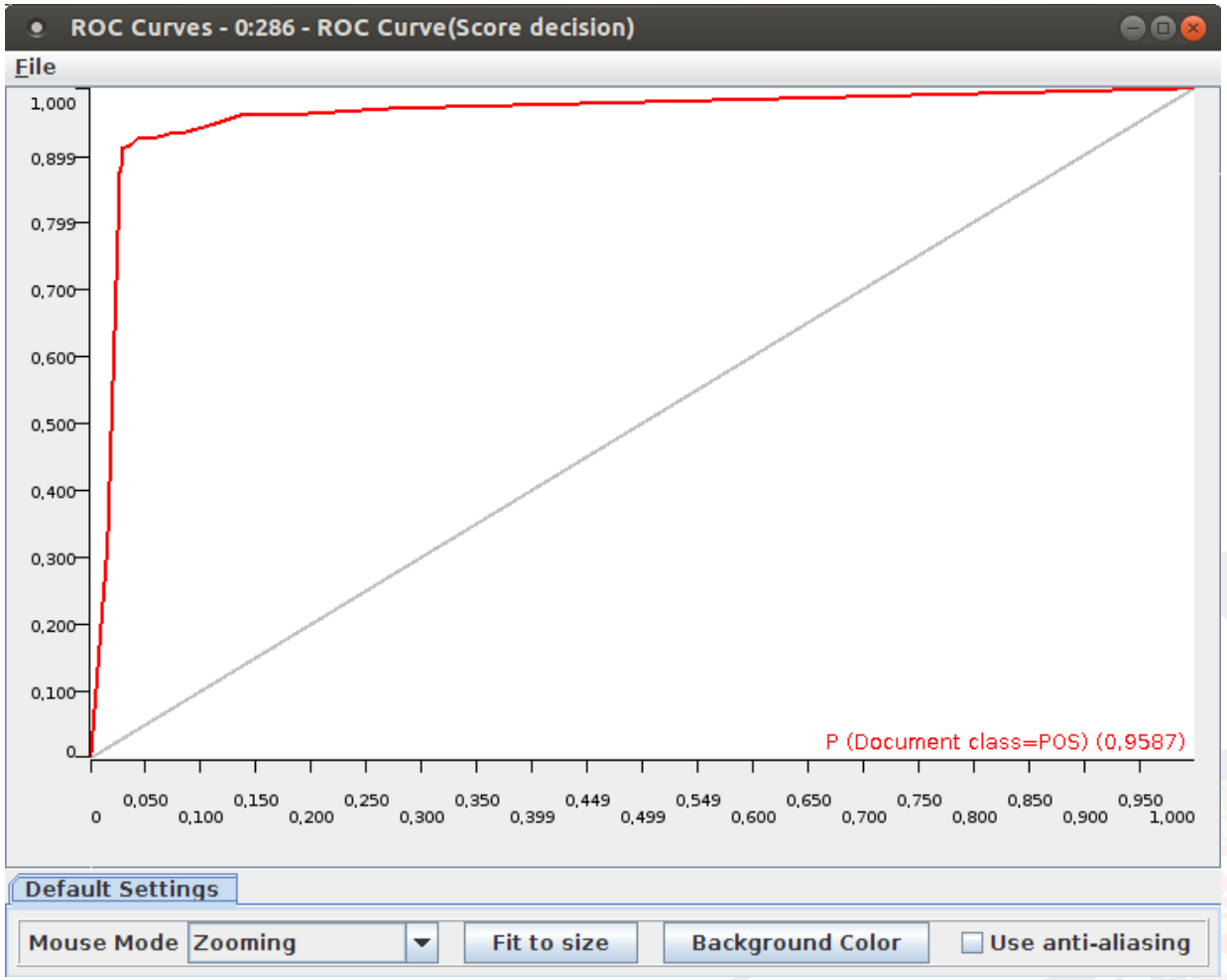
Podemos visualizar el árbol de decisión obtenido al realizar la clasificación. Existen dos vistas la normal y la simplificada

Visualización



Visualizamos los resultados mediante una curva ROC y una matriz de confusión, y se obtienen valores de *precisión*, *recall* y *F-Measure*.

Curva ROC



Visualizamos la curva ROC.

Ejercicio KNIME II

Crear un WorkFlow en KNIME que tomando los datos de IMDb, calcule la orientación del sentimiento del documento usando el MPQA Subjectivity Lexicon, y lo compare con los valores etiquetados, calculando la precisión de la solución empleada.

Repetir el mismo proceso empleando SentiWordNet 3.0 y SenticNet 5 (para ello será necesario descargar y leer los ficheros correspondientes, así como procesarlos adecuadamente).

Comparar los resultados obtenidos utilizando los diccionarios de sentimientos con los obtenidos usando dos de los clasificadores disponibles en KNIME.

Trabajo a Realizar

A entregar a través de PRADO (Fecha límite Lunes 15 Mayo):

- El proyecto KNIME

– Una breve documentación que debe contener:

  - Explicación de los pasos realizados para el procesamiento de los diccionarios de sentimientos SentiWordnet 3.0 y SenticNet 5.
  - Comparación y discusión de los resultados obtenidos con los tres diccionarios de sentimientos y dos técnicas de clasificación.

Evaluación Ejercicio KNIME

La calificación del Ejercicio será de 2 puntos de prácticas sobre el total de 3 puntos de prácticas del módulo II.