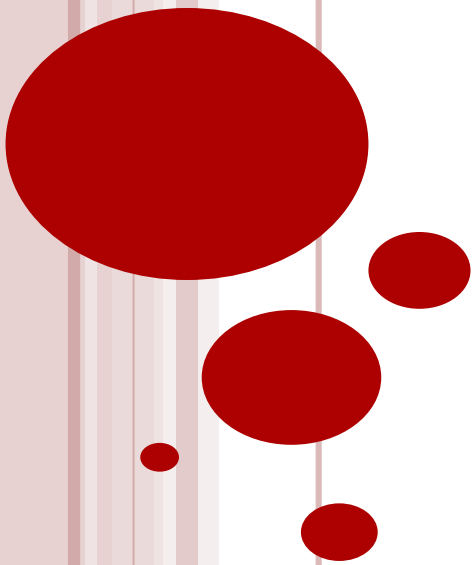
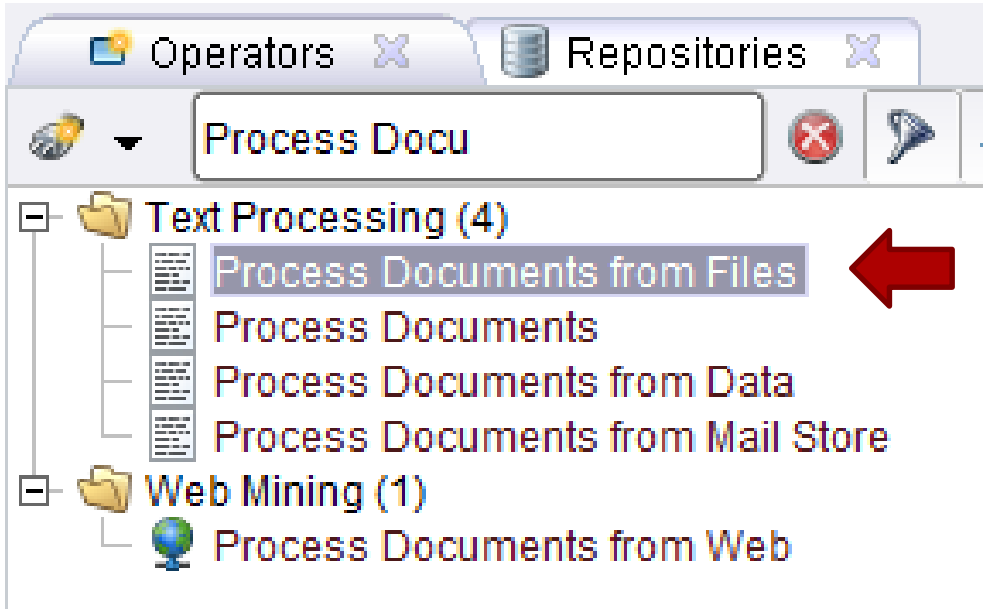


Clustering de documentos

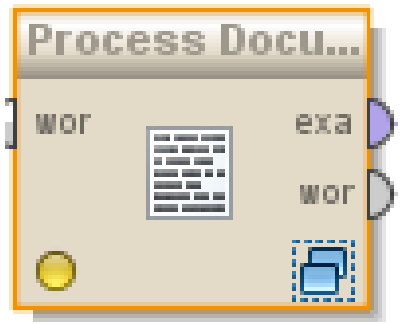
Clustering de documentos

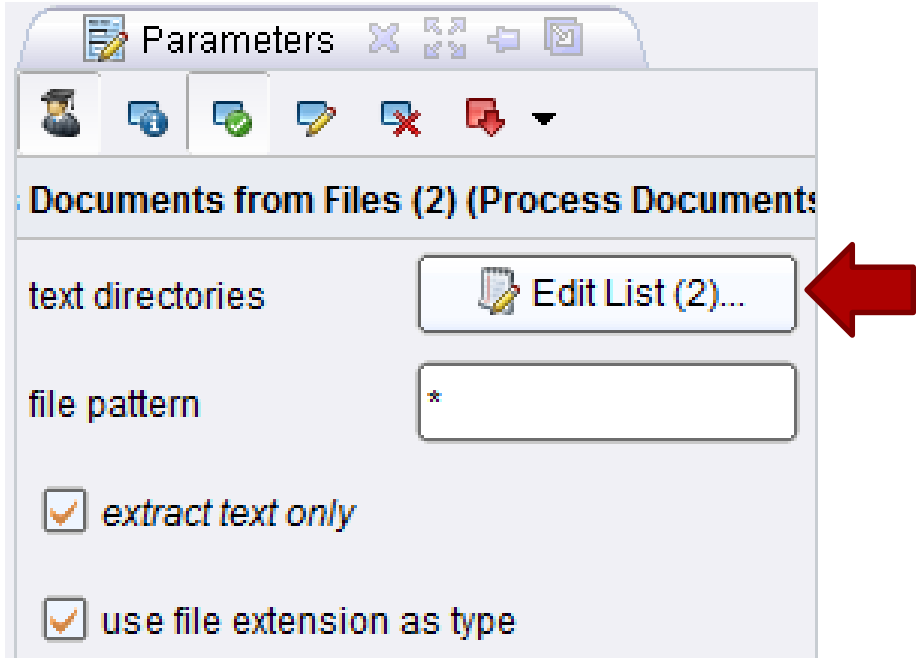
- Para aplicar una técnica de clustering basada en centroides es preciso tener una representación vectorial numérica de los datos de entrada.



Operador Process Documents from file

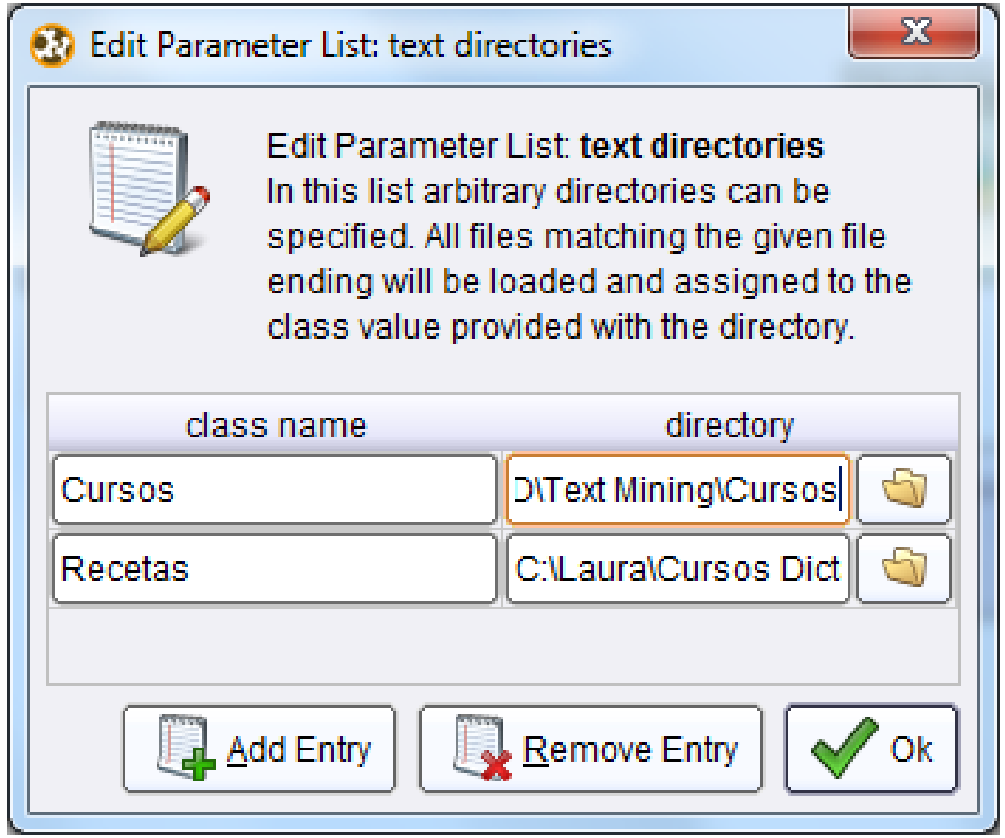
- Comencemos indicando la ubicación de los documentos





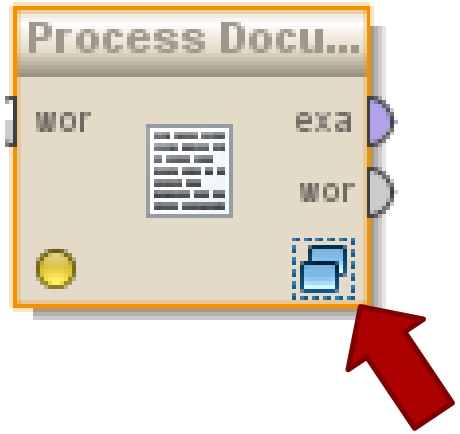
Operador Process Documents from file

- Ubicación de los documentos



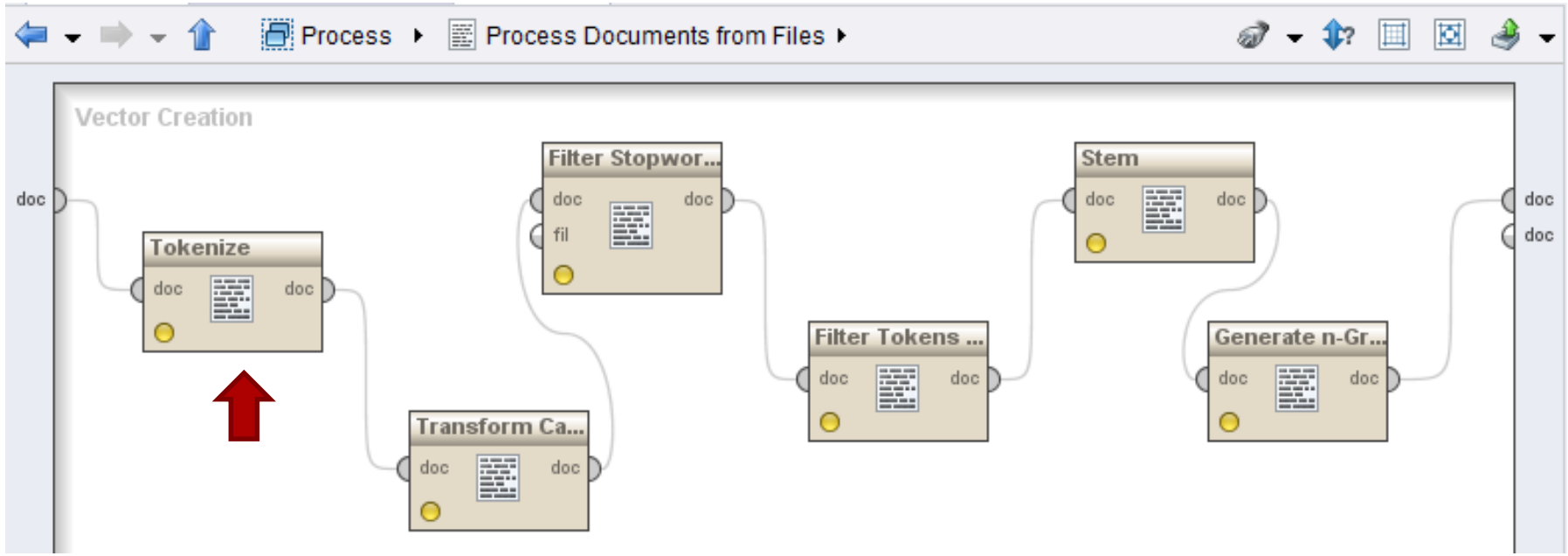
Operador Process Documents from file

- Dentro de este operador se indica la manera de separar cada texto en palabras



Cómo se obtienen los términos de los documentos

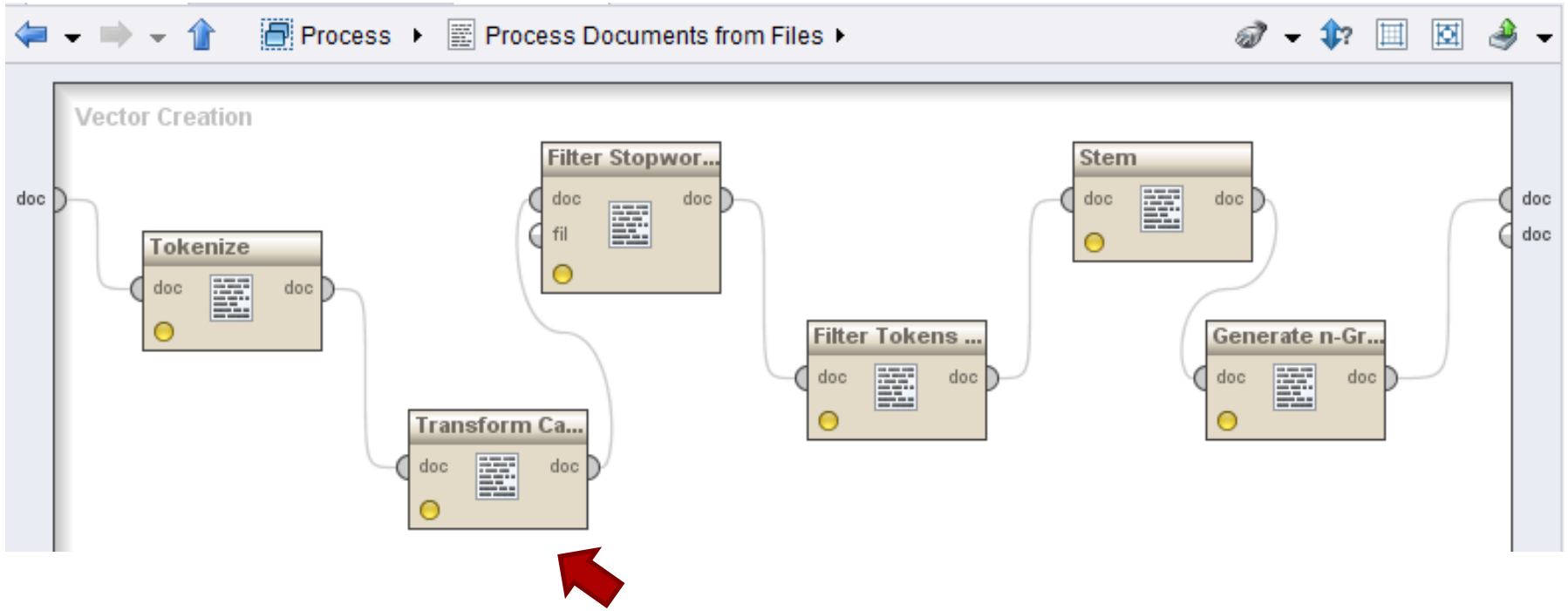
- Dentro del operador Process Documents from file



TOKENIZE
Separa cada documento en tokens
(puede indicarse el carácter a utilizar)

Cómo se obtienen los términos de los documentos

- Dentro del operador Process Documents from file

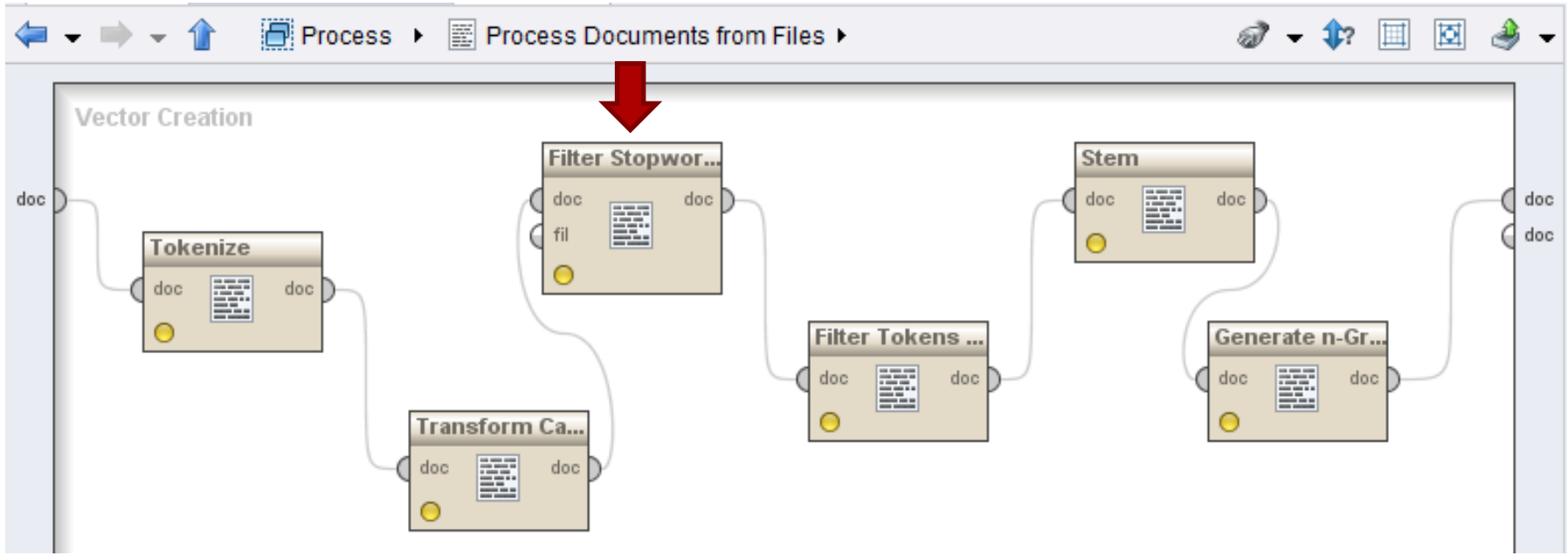


TRANSFORM CASES

Transforma todos los caracteres a minúsculas o mayúsculas según se indique (**seleccionar minúsculas**)

Cómo se obtienen los términos de los documentos

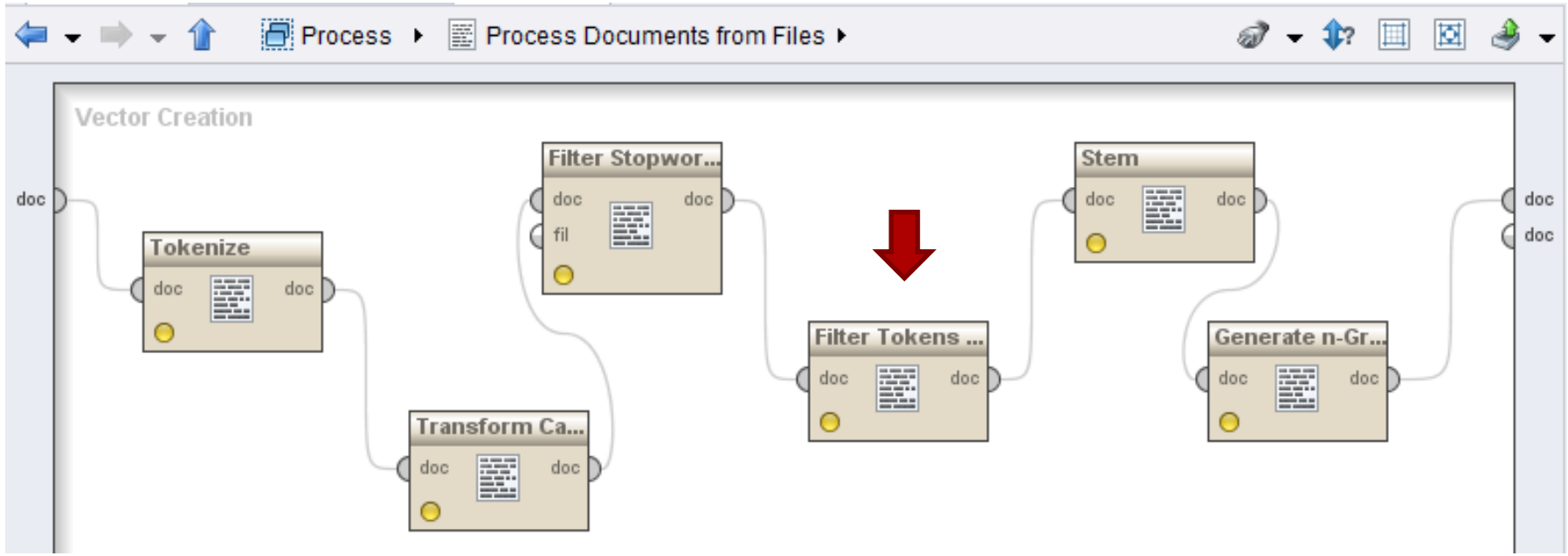
- Dentro del operador Process Documents from file



FILTER STOPWORDS (DICTIONARY)
Filtra los tokens que coincidan con cualquier stopword indicada en un determinado archivo
(utilice el archivo **stopwords_es.txt**)

Cómo se obtienen los términos de los documentos

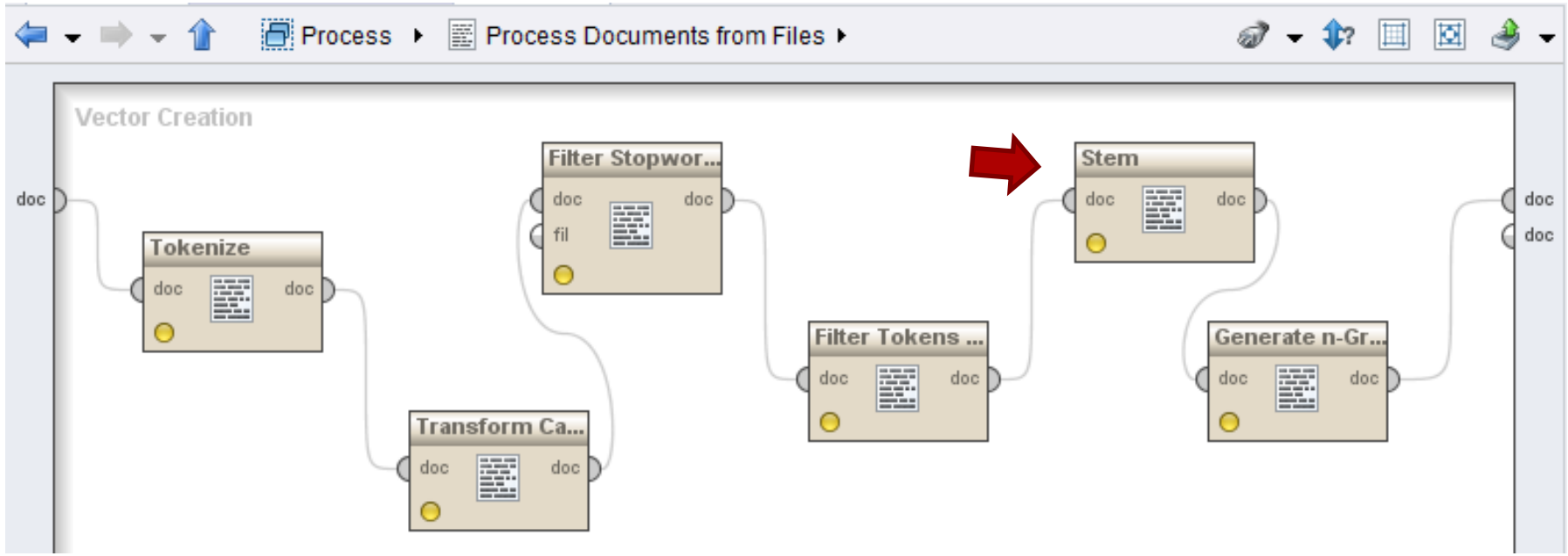
- Dentro del operador Process Documents from file



FILTER TOKENS (by LENGTH)
Configúrelo para que utilice sólo los tokens entre 4 (min chars) y 25 (max chars) caracteres.

Cómo se obtienen los términos de los documentos

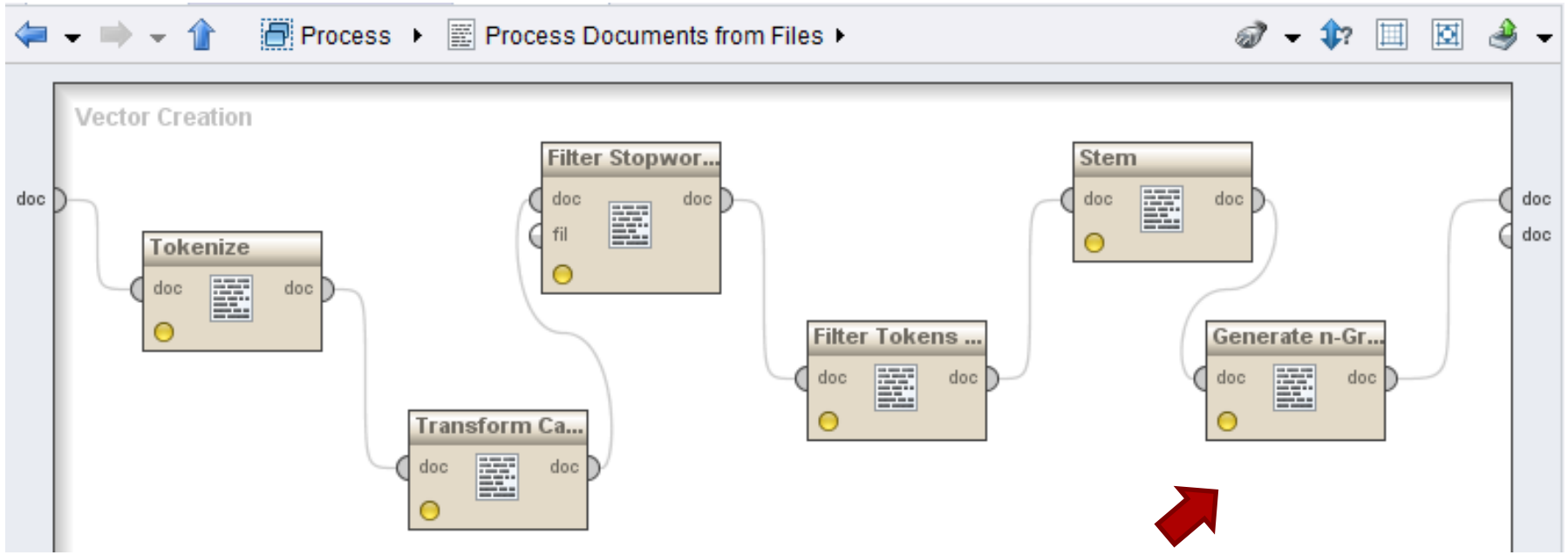
- Dentro del operador Process Documents from file



STEM (SNOWBALL)
Aplica un algoritmo de stemming para el lenguaje seleccionado.

Cómo se obtienen los términos de los documentos

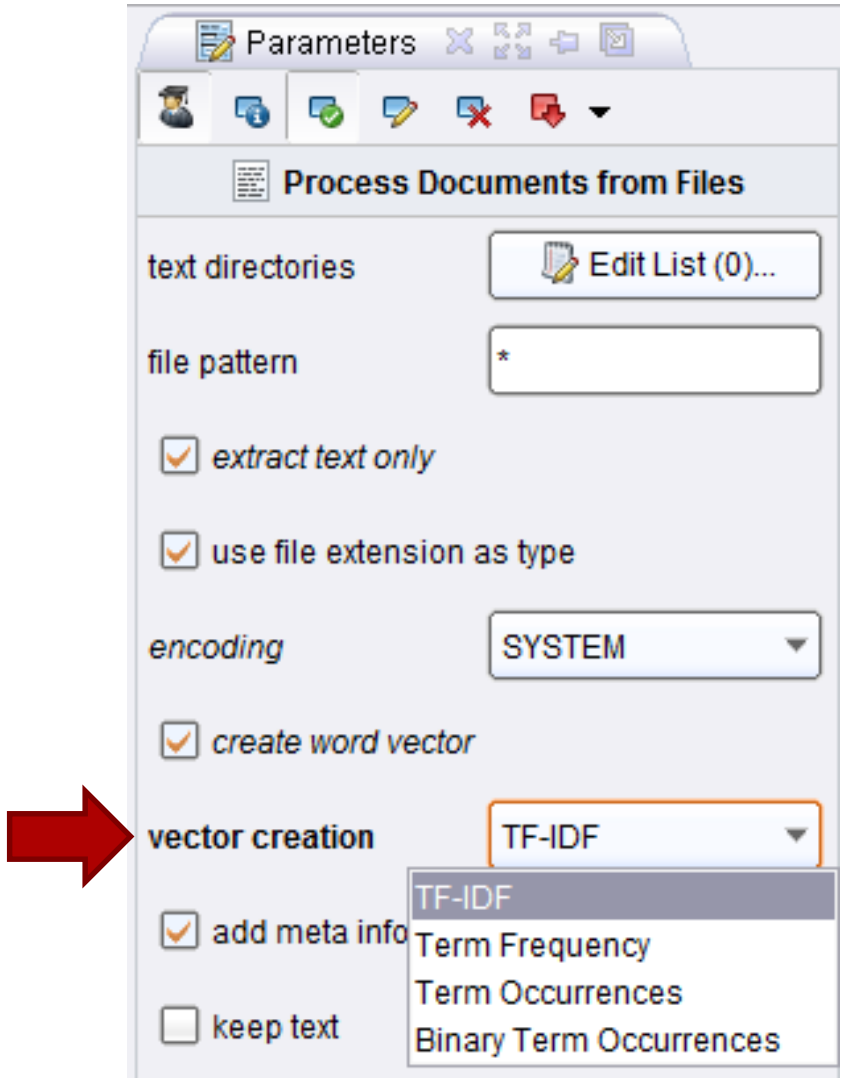
- Dentro del operador Process Documents from file



Grenerate n-Grams (terms)
Combina tokens. Utilice max length = 2

Representación numerica

- Se generará un gran diccionario con las palabras de todos los documentos.
- Es preciso indicar la representación numérica a utilizar



Frecuencia de un término

Opciones

- Si es binaria, $\boldsymbol{tf(t,d) = 1}$ si el término $\boldsymbol{t}$ aparece en el documento $\boldsymbol{d}$ y $\boldsymbol{0}$ en caso contrario.
- Si la escala es logarítmica

$$\boldsymbol{tf(t,d)} = \begin{cases} \boldsymbol{1 + log(\, f(t,d)\,)} & \boldsymbol{f(t,d) > 0} \\ \boldsymbol{0} & \boldsymbol{f(t,d) = 0} \end{cases}$$

- La frecuencia de un término puede normalizarse utilizando

$$\boldsymbol{tf(f,d)} = \frac{\boldsymbol{f(t,d)}}{\boldsymbol{max\{f(t,d) : t \in d\}}}$$

INVERSE DOCUMENT FREQUENCY

- Es una medida de cuán común o raro es un término entre todos los documentos

$$\boldsymbol{idf}(\boldsymbol{t}, \boldsymbol{D}) = \log\left(\frac{|\boldsymbol{D}|}{|\{\, \boldsymbol{d} \in \boldsymbol{D} : \boldsymbol{t} \in \boldsymbol{d}\}|}\right)$$

siendo

- $|\boldsymbol{D}|$ el número total de documentos.
- $|\{\, \boldsymbol{d} \in \boldsymbol{D} : \boldsymbol{t} \in \boldsymbol{d}\}|$ el número de documentos en los que aparece el término $\boldsymbol{t}$. Puede evitar la división por 0 usando $\mathbf{1} + |\{\, \boldsymbol{d} \in \boldsymbol{D} : \boldsymbol{t} \in \boldsymbol{d}\}|$
- La base del logaritmo es sólo un factor constante multiplicativo.

TF-IDF

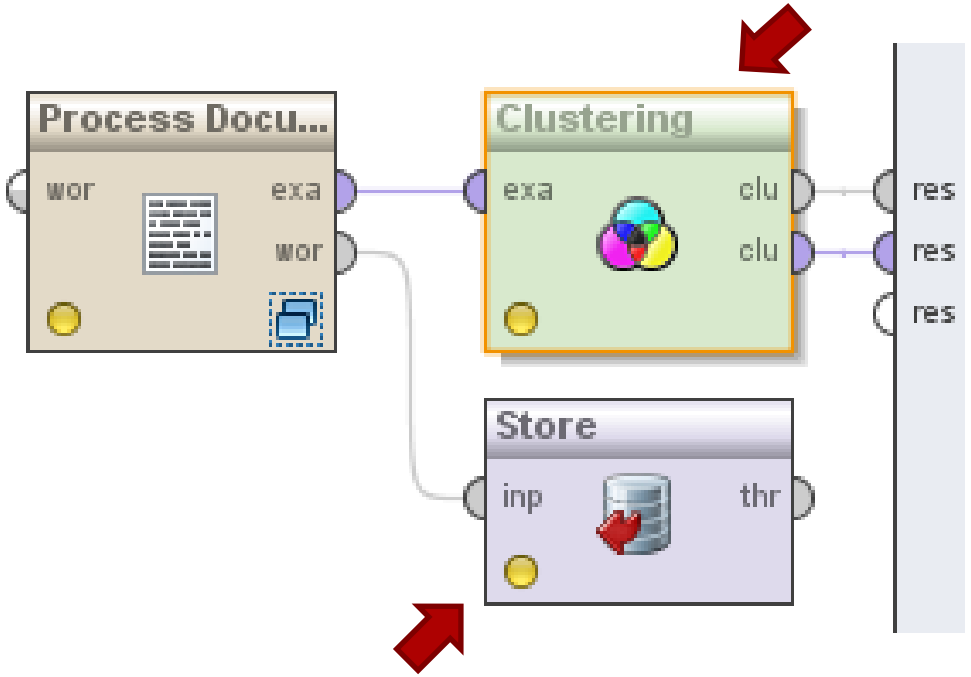
- Se calcula de la siguiente forma

$$\boldsymbol{tf - idf(t, d, D) = tf(t, d) * idf(d, D)}$$

  - Se obtendrán valores altos en aquellos términos que posean frecuencia alta (en el documento dado) con una frecuencia baja del término en la colección completa de documentos, filtrando de esta forma los términos comunes.
  - Dado que la proporción usada como argumento de la función $\boldsymbol{log}$ en $\boldsymbol{idf}$ es $> \mathbf{0}, \boldsymbol{tf - idf}$ también es $> \mathbf{0}$ siempre.
  - Un término que aparece en muchos documentos hace que la proporción del logaritmo en $\boldsymbol{idf}$ se acerque a $\mathbf{1}$ y por lo tanto $\boldsymbol{idf}$ y $\boldsymbol{tf - idf}$ tenderán a $\mathbf{0}$.

AGRUPAMIENTO

Clustering (k-Means)
Use k=3



Ejecutar y observar los resultados obtenidos

STORE
Almacena las palabras en el repositorio