23 Junio 2020. Text Mining. Weka. Parte 1

Construya un tutorial en que detalle el desarrollo de la tarea.

Archivo de datos de ejemplo de weka: ReutersGrain-train.arff y ReutersGrain-test.arff

Los datos contienen el texto de artículos, si el atributo se refiere a grain entonces class es 1, si no se refiere class es cero.

El objetivo es que a partir de los datos se construye un clasificador de artículos, para identificar si hablan o no de granos (grain)

Los archivos forma parte de los datos de ejemplo de weka, se encuentran en la carpeta de instalación de weka

(dentro de data).

Abrir el archivo ReutersGrain-train.arff con el bloc de notas para poder visualizar los datos contenidos y cerrar sin guardar.

1. Abrir en Weka el archivo ReutersGrain-train.arff. El archivo forma parte de los datos de ejemplo de weka, se encuentra en la carpeta de instalación de weka

(dentro de data).

2. El archivo tiene dos atributos: Text y class. ¿Cuantos artículos corresponden a casa clase? Se refiere a grain , no se refiere a grain

3. En classify, text options, segunda opción se selecciona Supplied test test, clic en Set se carga el archivo: ReutersGrain-test.arff

¿Qué hicimos al seleccionar esta opción?

4. Para clasificar se construye un árbol de decisión. Pero en el caso de datos textuales, se define un clasficador filtrado, que preprocesará los datos

con el filtro seleccionado antes de construir el árbol de decisión.

5. En Classify seleccionar meta y luego FilteredClassifier.

6. A la derecha de Choose nos aparece la posibilidad de configurar el filtro como sigue:

7. En classifiers seleccionar trees y J48

8. Clic abajo en Filter

9. Seleccionar unsupervised

10. Seleccionar atribute

11. Seleccionar StringToWord vector (no modificar las opciones por defecto)

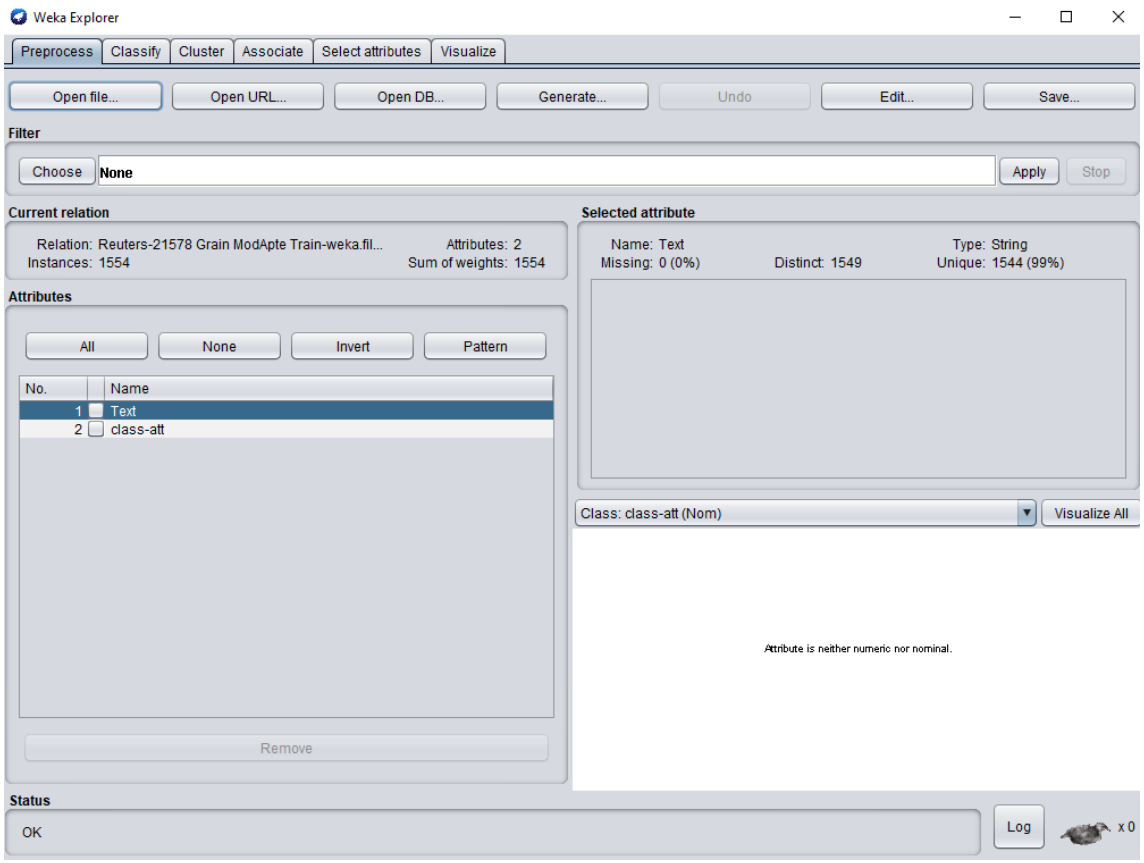
12. Start para ejecutar

13. Presentar y comentar la matriz de confusión

14. Mostrar el árbol de decisión generado.

1.

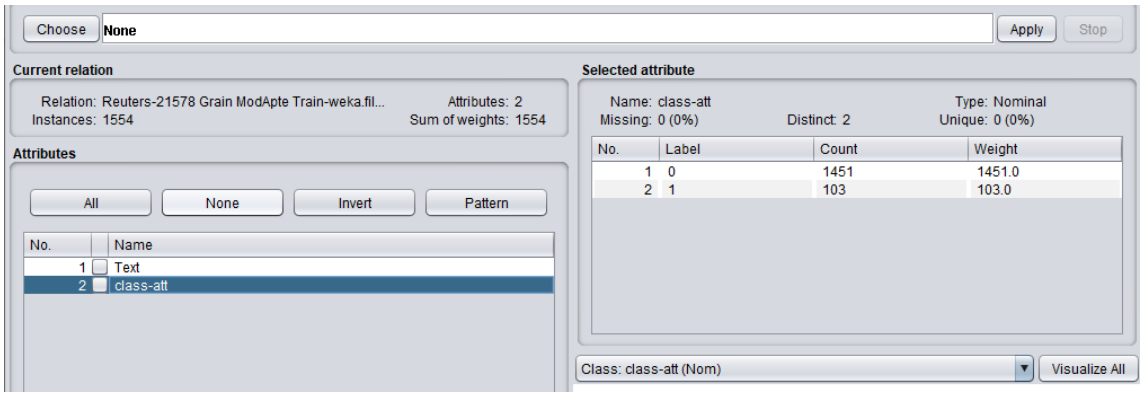
Primeramente, abrimos la herramienta Weka. Hacemos click en Explorer y se abrirá una nueva ventana. En esta ventana abriremos el archivo con el que vamos a trabajar haciendo click en Open File. El archivo es un archivo propio de Weka que trae como ejemplo. Estos se encuentran en la carpeta que se crea al instalar Weka.



2.

A la clase se refiere a grain, podemos observar que le corresponden 103 artículos.

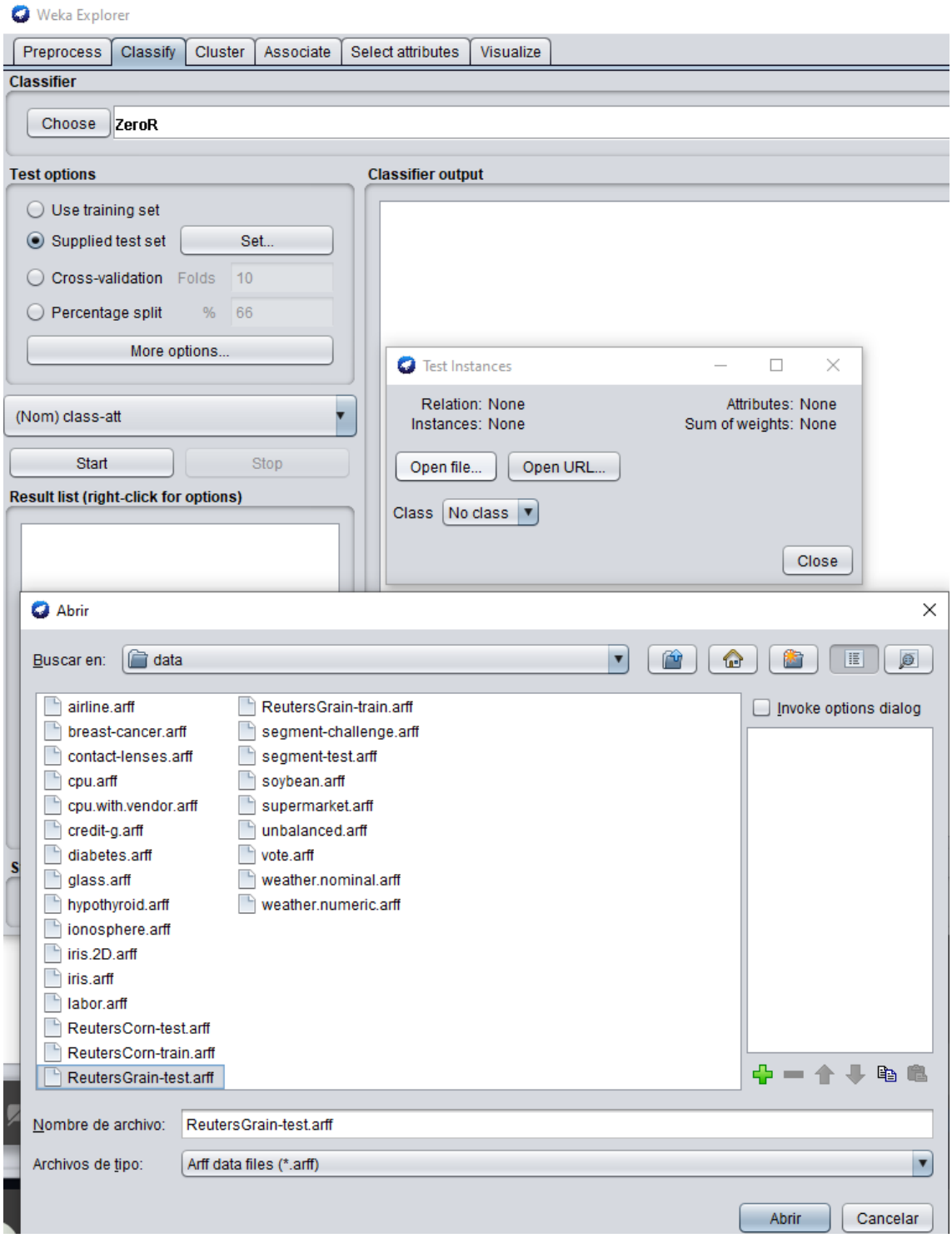
A la clase no se refiere a grain, podemos observar que le corresponder 1451 artículosç.



3.

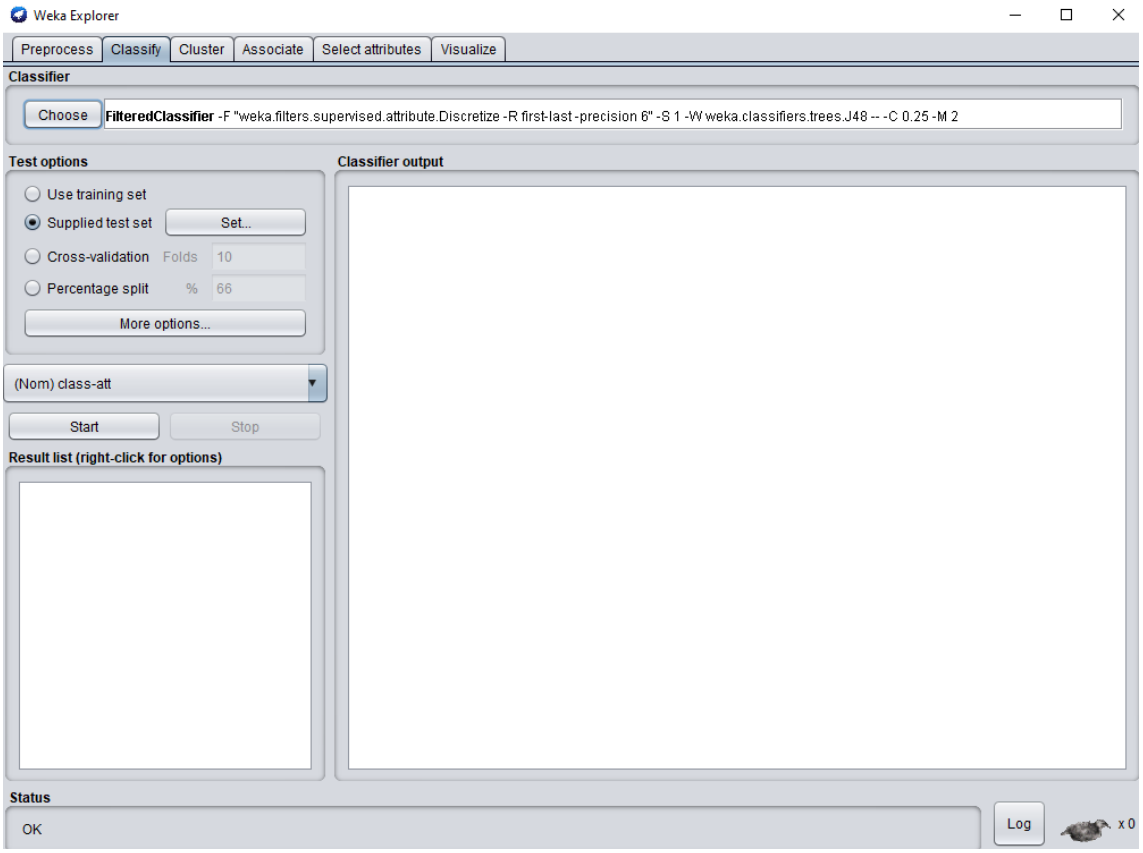
En la pestaña Clasify, seleccionamos Set, que se encuentran al lado de Supplied test set, y seleccionamos Open File y seleccionaremos el archivo que contiene los datos de testeo (ReutersGrain-test.arff).

De esta manera, lo que hacemos es proporcionar un archivo con datos para testear el modelo.

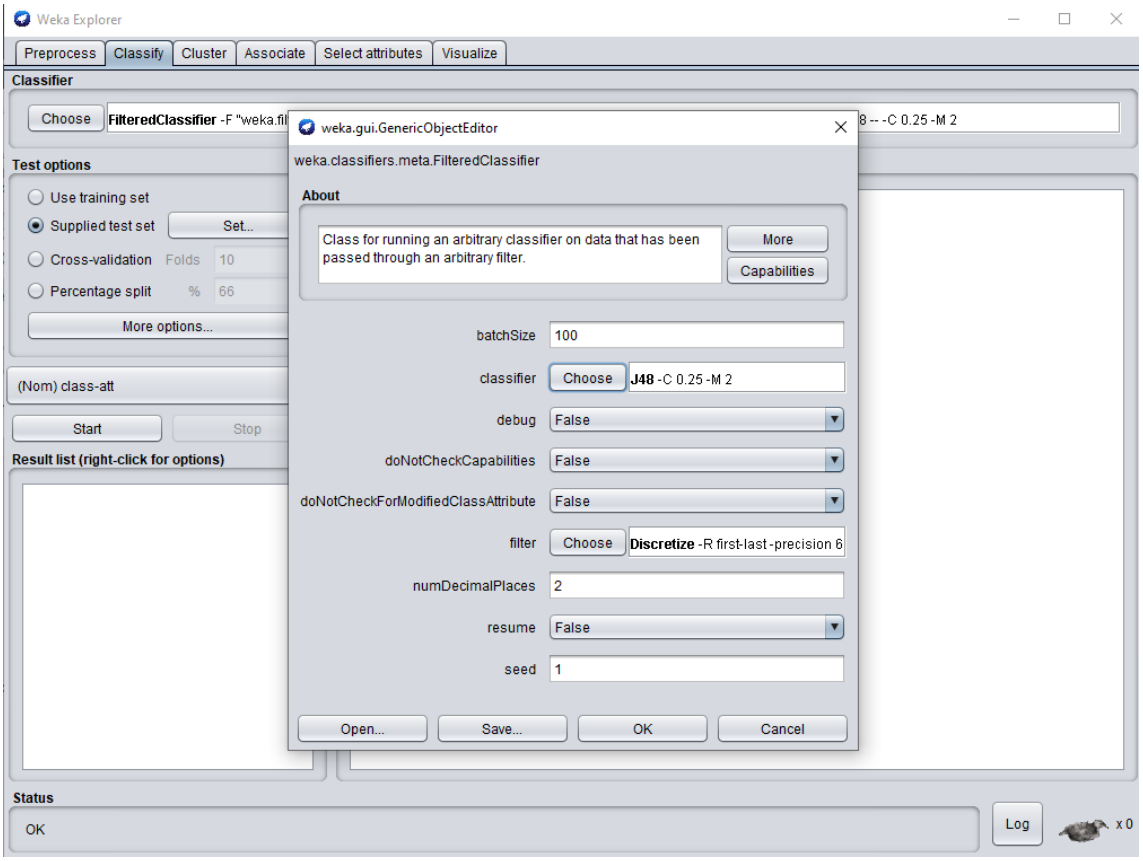


4, 5 y 6.

Primero vamos a preprocesar los datos. Una opción de filtrado sería seleccionar en Clasify→Choose y seleccionamos en la carpeta Meta→FilteredClassifier



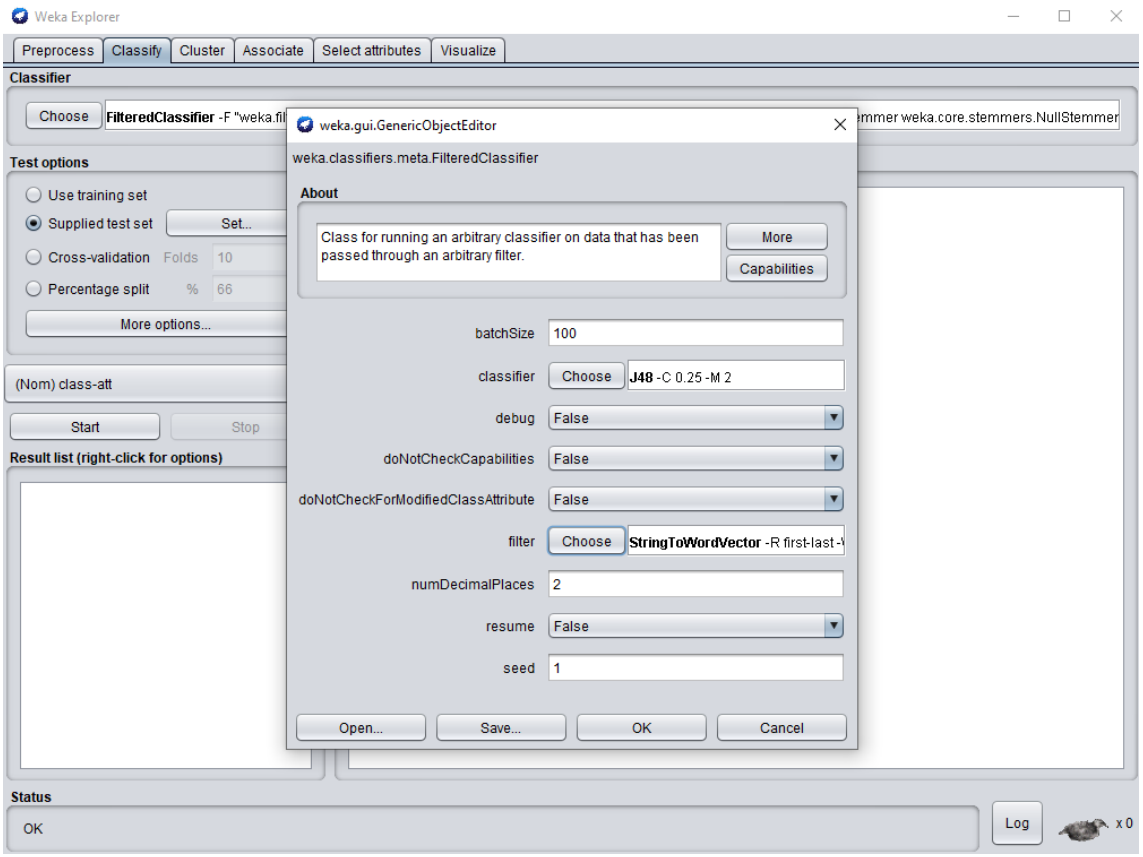
Haciendo click en FilteredClassfier podemos configurar el filtro.



Seleccionamos en classifier→Choose→J48

8, 9, 10 y 11.

Luego seleccionamos filter→Choose→Unsupervisated→Atribute→StringToWord

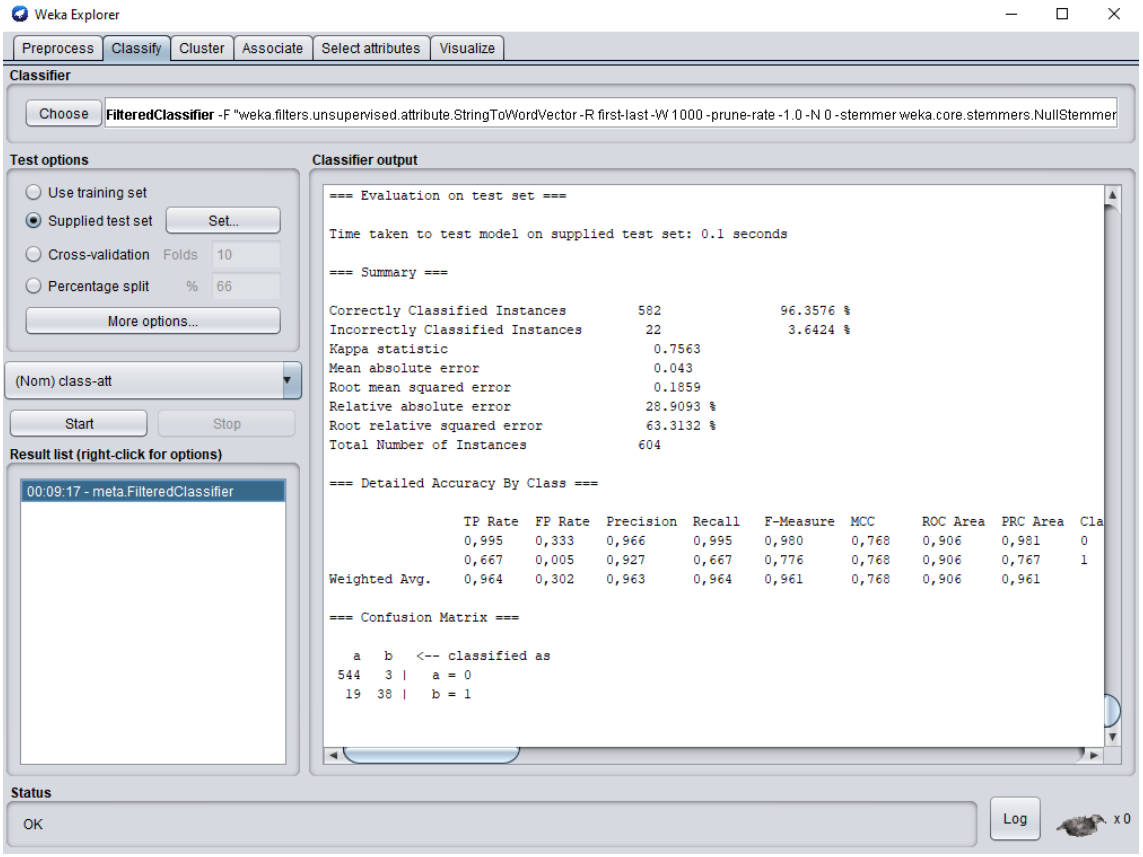


Esto lo hacemos para convertir cada palabra en un atributo.

Tras esto, ya estamos en disposición de ejecutar.

12, 13 y 14.

Pulsamos Star para ejecutar el algoritmo. Se nos abrirá una ventana que presentará los datos recabados por el mismo. Entre ellos, podremos observar la matriz de confusión.



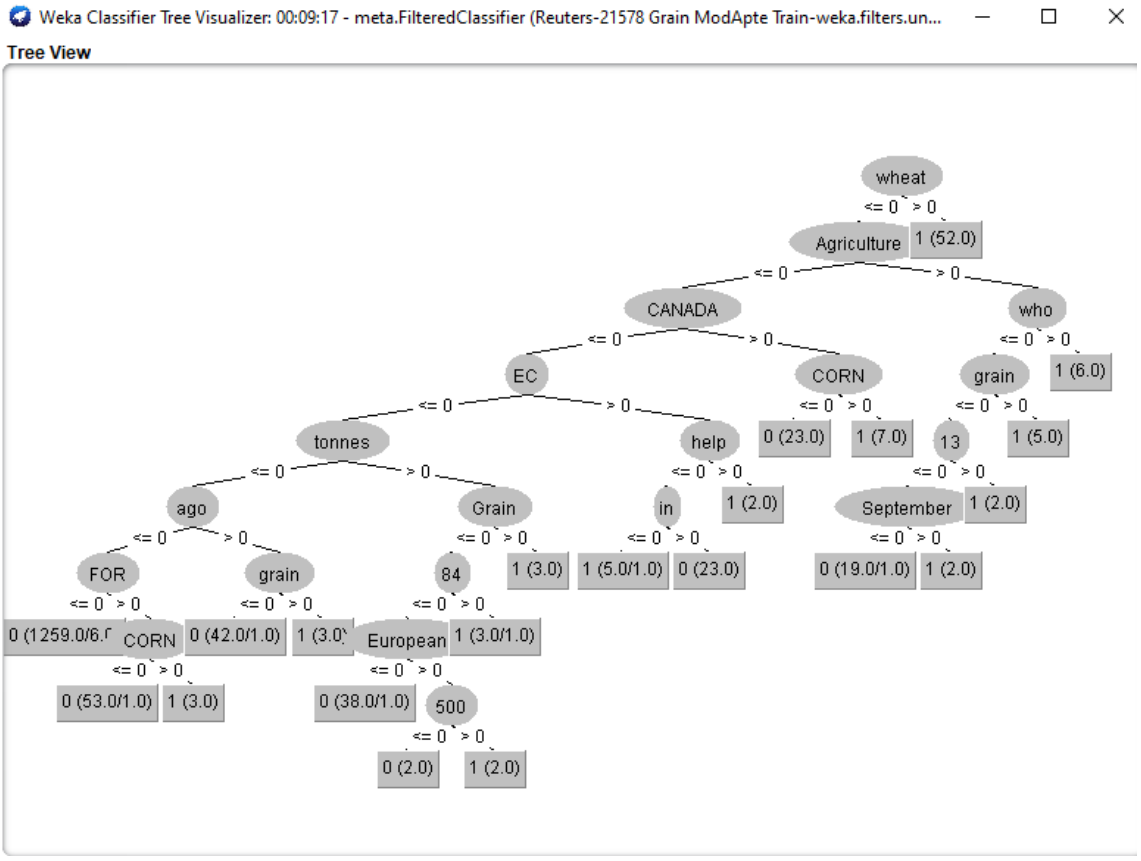
Podemos ver a raíz de la matriz de confusión, que el algoritmo clasificó bastante bien los datos, ya que solamente 19 elementos que implicaban no grain se clasificaron en grain, y 3 elementos que implicaban grain se clasificaron como no grain. Sin embargo, podemos observar que un total de 582 elementos se clasificaron correctamente.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 544   3 |   a = 0
  19  38 |   b = 1
```

Hacemos click derecho sobre Result List→meta.FilterdClassifier y seleccionamos visualize tree.

Los nodos vemos que son palabras, ya que estamos trabajando con textos.


Weka Classifier Tree Visualizer: 00:09:17 - meta.FilteredClassifier (Reuters-21578 Grain ModApte Train-weka.filters.un...
Tree View
wheat
<= 0
> 0
Agriculture
1 (52.0)
<= 0
> 0
CANADA
who
<= 0
> 0
<= 0
> 0
EC
CORN
grain
1 (6.0)
<= 0
> 0
<= 0
> 0
<= 0
> 0
tonnes
help
0 (23.0)
1 (7.0)
13
1 (5.0)
<= 0
> 0
<= 0
> 0
<= 0
> 0
ago
Grain
in
1 (2.0)
September
1 (2.0)
<= 0
> 0
<= 0
> 0
<= 0
> 0
FOR
grain
84
1 (3.0)
1 (5.0/1.0)
0 (23.0)
0 (19.0/1.0)
1 (2.0)
<= 0
> 0
<= 0
> 0
<= 0
> 0
0 (1259.0/6.
CORN
0 (42.0/1.0)
1 (3.0
European
1 (3.0/1.0)
<= 0
> 0
<= 0
> 0
0 (53.0/1.0)
1 (3.0)
0 (38.0/1.0)
500
<= 0
> 0
0 (2.0)
1 (2.0)